RB-68-38

SOME TEST THEORY FOR TAILORED TESTING

Frederic M. Lord

Office of Naval Research Contract Nonr-2752(00)
Project Designation NR 151-201
Frederic M. Lord, Principal Investigator

Educational Testing Service
Princeton, New Jersey

September 1968

SOME TEST THEORY FOR TAILORED TESTING

Frederic M. Lord

Office of Naval Research Contract Nonr-2752(00)
Project Designation NR 151-201
Frederic M. Lord, Principal Investigator

Educational Testing Service
Princeton, New Jersey

September 1968

SOME TEST THEORY FOR TAILORED TESTING

Abstract

In a tailored test, each item is selected for administration on the basis of the examinee's responses to previous items, with a view towards optimum measurement of this particular examinee. Various simple rules for 1) selecting the items to be administered and 2) scoring the examinee's responses are compared and evaluated. Some fundamental ideas emerge that will serve as guides in the future design of tailored testing programs.

ADDENDUM TO ETS RB-68-38

A practical method of evaluating Robbins-Monro procedures for tailored tests, without using Monte Carlo methods, can be adapted from a method of W. G. Cochran and M. Davis described in "The Robbins-Monro method for estimating the median lethal dose", Journal of the Royal Statistical Society, Series B (Methodological), 1965, 27, 28-44. Investigations using this method are under way.

Table of Contents

SOME TEST THEORY FOR TAILORED TESTING*

It seems likely that in the not too distant future, many mental tests will be administered and scored by computer. Computerized instruction will be common, and it will be convenient to use computers to administer achievement tests also (Turnbull, 1968).

The computer can test many examinees simultaneously, with the same or with different tests. If desired, each examinee can be allowed to answer test questions at his own rate of speed. This situation opens up new possibilities. The computer can do more than simply administer a predetermined set of test items. Given a pool of precalibrated items to choose from, the computer can design a different test for each individual examinee.

An examinee is measured most effectively when the test items are neither too hard nor too easy for him. Thus, for any given psychological trait, the computer's main task at each step of the test administration might be to estimate tentatively the examinee's level on the trait, on the basis of his responses to whatever items have already been administered. The computer could then choose the next item to be administered on the basis of this tentative estimate.

Such testing has been called "branched testing," "programmed testing," "sequential item testing," and "computerized testing." Clearly, the procedure could be implemented without a computer. Here, emphasizing the key feature, we will speak of tailored testing.

---

*This work was supported in part by contract Nonr-2752(00) between the Office of Naval Research and Educational Testing Service. Reproduction, translation, use and disposal in part by or for the United States Government is permitted.

It should be clear that there are important differences between testing for instructional purposes and testing for measurement purposes. The virtue of an instructional test lies ultimately in its effectiveness in changing the examinee. At the end, we would like him to be able to answer every test item correctly. A measuring instrument, on the other hand, should not alter the trait being measured. Moreover, as already noted, measurement is most effective when the examinee knows the answers to only about half of the test items. The discussion here will be concerned exclusively with measurement problems and not at all with instructional testing.

Sections 3-6 contain necessary technical preliminaries for formulating and dealing with the problem. Section 8 discusses key questions in evaluating different testing procedures. Sections 7, 10, 12, 17, and 19 derive and present mathematical formulas necessary for describing and evaluating various tailored testing procedures. Sections 9 and 12-19 present some of the numerical results obtained for various testing procedures in various situations. Sections 2 and 11 are devoted to general discussion. A partial summary is given in section 20.

It is a fortunate fact that most of the problems dealt with here closely parallel similar problems in bioassay. Much fruitful work has been done on the bioassay problems. This provides the inspiration, the background, and indeed the backbone of the present report. A brief discussion of this bioassay work is given in section 11 at a point where the similarities and differences with tailored testing problems can be discussed intelligibly.

1. A Statement of the Problem

When the frequency distribution of the relevant psychological trait in the group to be tested is well known from previous testings of similar

groups, a Bayesian analysis, using group statistics, is appropriate. Such an analysis would reach different conclusions depending on the frequency distribution of the trait. The present exploratory treatment will not use Bayesian analysis. Here we will be concerned throughout with the problem of "measuring" a single examinee with respect to one psychological dimension. Since each examinee is to be considered by himself, group statistics (for example, test reliability coefficients) will play only a very marginal role.

The notion of "measuring" an examinee implies that there is some numerical value θ, say, characterizing him, which we wish to determine or estimate. The data available for making this estimate will be the examinee's responses to whatever test items are administered to him. The basic problem is to choose n test items for administration so that his n responses will enable us to estimate θ as efficiently as possible.

The optimum set of n items for this purpose depends on the unknown value of θ. Because of this fact, it is not clear that an optimum strategy exists, independent of the unknown θ, for choosing the desired n items. In any case, we will not even attempt here to find an optimum strategy. Instead, we will try to evaluate certain available simple strategies with a view to learning which of these are superior to others and what considerations seem relevant for determining their various virtues.

2. Some Strategies

Current research in tailored testing (see Linn, Rock, & Cleary, 1969 for references; also Hansen & Schwarz, 1968) is typically built on the following rule. If the examinee answers an item correctly, the next item administered should be harder; if he answers it incorrectly, the next item should be easier. This will be referred to as the branching rule. An obvious question

to be answered here is how much the item difficulty should be varied from item to item. A second question of strategy is how to score the responses once the items have been administered. Various scoring methods have been tried, as will be seen.

Suppose for the moment that n, the number of items to be administered, is indefinitely large and that the branching rule is used. Suppose further that at the start large differences in difficulty are used from item to item and that these differences are gradually reduced until ultimately successive items are of nearly equal difficulty. Will such a strategy allow us to pinpoint the item difficulty level at which the examinee answers exactly half the items correctly, in the long run? If so, we can characterize the examinee's ability level in terms of this difficulty level.

The process just described is a Robbins-Monro process (Robbins & Monro, 1951). Conditions for its convergence to the desired value are not difficult to satisfy in practice. The entire process will be discussed in section 19.

The point to be made here is that the practical use of the branching process to estimate the examinee's ability does not require strongly restrictive assumptions. It is not necessary for this purpose to know the exact mathematical form of the dependence of item response on examinee ability θ and on parameters, such as item difficulty, describing the item.

If we wish to evaluate and compare the efficiency of different methods for estimating examinee ability, however, it becomes necessary to have some further information. In any one particular case, this information could be gathered by exhaustive testing of the particular examinee, provided this testing could be done without changing him in the process. For purposes of the present paper, in order to generalize our conclusions to as yet untested populations of examinees, we will instead make assumptions about the characteristic curves of the test items.

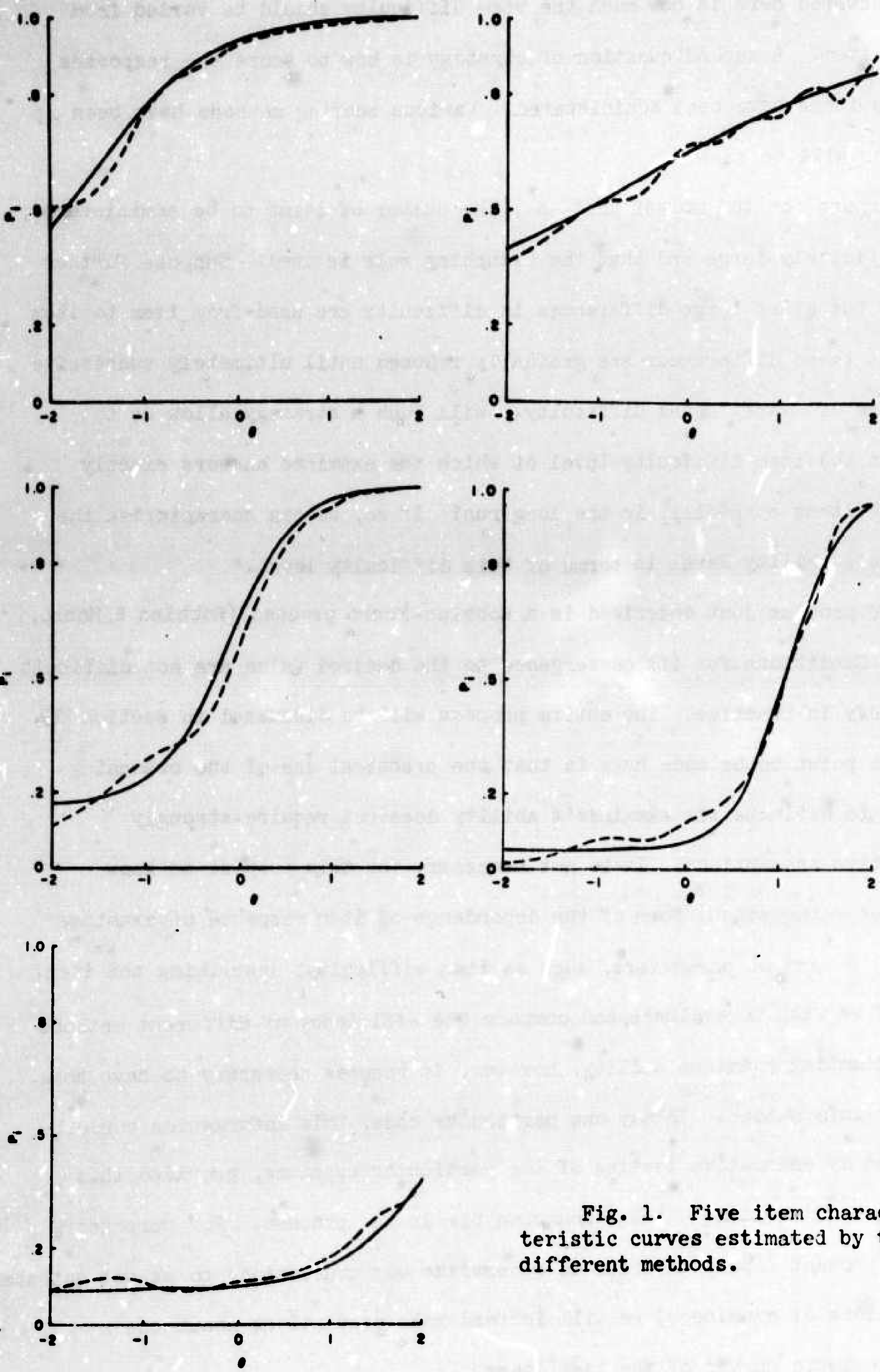


Fig. 1. Five item characteristic curves estimated by two different methods.

3. Item Characteristic Curves

An item characteristic curve represents the probability of a correct answer to an item as a function of the trait θ being measured. If the item is scored zero or one, this curve is automatically also the regression of item score on θ. Icc's are important, first of all, because they enable us to quantify important characteristics of individual test items. Secondly, because they enable us to predict, probabilistically, how the examinee will respond to any chosen item.

Estimated characteristic curves are shown in Figure 1 (reproduced from Lord, 1968a) for five actual test items. All these curves have the typical ogive shape with an upper asymptote at P_i (probability of a correct answer) equals 1.0 and a lower asymptote at $P_i = c_i$, $0 \leq c_i \leq 1$, where c_i is a parameter characterizing item i.

The solid curves shown are all logistic functions. When $c_i = 0$, the logistic function is simply

$$P_i == P_i(\theta) = \frac{1}{1 + \exp\left[-1.7a_i(\theta - b_i)\right]}, \quad (-\infty < \theta < \infty), \tag{1}$$

where a_i and b_i are parameters describing the test item. (The symbol == is used here and elsewhere to indicate a definition.)

When test items can be answered correctly by random guessing, then $P_i > 0$ for all θ and $c_i > 0$. In this case, we sometimes use the three-parameter logistic function

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp\left[-1.7a_i(\theta - b_i)\right]} , \quad (-\infty < \theta < \infty) . \tag{2}$$

This is the function shown by the solid curves in Figure 1. Logistic icc's are discussed in detail by Birnbaum (1968).

The logistic function in (1) is in some ways a close approximation to the <u>normal ogive</u>

$$P_i(\theta) = \Phi[a_i(\theta - b_i)] = \int_{-\infty}^{a_i(\theta - b_i)} \phi(u)\, du , \quad (-\infty < \theta < \infty) , \tag{3}$$

where Φ is defined by (3) and

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} u^2\right) . \tag{4}$$

(1) and (3) do not differ by as much as .01 for any value of θ (the ratio of (1) to (3) is large in the tails, however). When guessing occurs, (3) is replaced by

$$P_i(\theta) = c_i + (1 - c_i)\, \Phi[a_i(\theta - b_i)] , \quad (-\infty < \theta < \infty) . \tag{5}$$

We assume here that (1), (2), (3), or (5) holds true for a given examinee regardless of any knowledge that may be available about his performance on items other than item i. This means that when θ is fixed, the probability of the examinee answering a fixed set of n items correctly is simply $\prod_{i=1}^{n} P_i(\theta)$, the product of the separate probabilities.

A common question is whether it may be possible by empirical studies to establish the superiority of either the logistic or the normal-ogive model for icc's. The appropriate answer to this question is probably that if an

empirical study could be made sensitive enough to discriminate between these two models, it would almost surely be sensitive enough to prove that neither model was strictly correct.

Except where otherwise noted, the present paper assumes that all icc's are of the form (3) or (5). Similar assumptions have proven to be very valuable in bioassay work. The reader who feels uncomfortable with such assumptions should consider the report by Lord (1968), in which (2) was found to agree closely with other estimates of icc's obtained without prior assumption regarding their mathematical form. These latter estimates are shown in Figure 1 by the curved dashed lines.

4. Item Parameters

The parameters a_i, b_i, and c_i will be used to describe items. In this report, we are primarily concerned with the problem of selecting items to be administered. This will be done on the basis of their parameters, determined in advance by pretesting. Thus it will be worthwhile here to examine the meaning of these parameters.

As already noted, c_i determines the lower asymptote. Accordingly, $0 \leq c_i \leq 1$. Any examinee, however low his θ, has a chance $> c_i$ of answering the item correctly. Thus, c_i will here be thought of as the probability of chance success on item i as a result of random guessing.

All of the curves (1), (2), (3), and (5) have an inflexion point, which is also a center of symmetry for the curve. The parameter b_i gives the abscissa of the inflexion point, which is at $\theta = b_i$. We will allow b_i to assume any value in the range $-\infty < b_i < \infty$. It is clear that b_i represents the <u>difficulty</u> of the item. The larger the value of b_i, the less likely an examinee is to answer the item correctly.

The value a_i is closely related to the slope of the icc at the inflexion point. For the three-parameter logistic, this slope is $.425(1 - c_i)a_i$; for the three-parameter normal ogive, it is $.3989(1 - c_i)a_i$. The parameter a_i is spoken of as representing the <u>discriminating power</u> of the item. For $a_i \geq 0$, the larger the value of a_i, the more the item discriminates between examinees with high θ and examinees with low θ. We plan to use items that are positively correlated with θ; consequently we shall restrict a_i to the range $0 < a_i < \infty$.

In this report, we assume that all items have been calibrated -- the item parameters estimated -- by pretesting in advance of any use we make of them. How this is done is not our problem here, but we will glance at it briefly. Presumably, over a period of years or decades, a large pool of items with accurately estimated parameters will gradually be accumulated.

The item parameters apparently can be estimated by maximum likelihood (Lord, 1968b; Bock, 1967; Birnbaum, 1968, section 17.9). This is at present a costly and hazardous operation. The parameters can be approximated by more familiar procedures, which we mention below. These approximations are based on the assumption of normal ogive icc's with $c_i = 0$ and on the rather unlikely assumption that θ is normally distributed in the group tested. Since the unit of measurement used to express θ is arbitrary, we will choose it so that $\sigma_\theta = 1$.

Under the assumptions stated, the parameter a_i can be estimated with the help of the relation (Lord & Novick, 1968, section 16.10):

$$a_i = \frac{\rho'_{\theta i}}{\sqrt{1 - \rho'^2_{\theta i}}} , \qquad (6)$$

where $\rho'_{\theta i}$ is the biserial correlation of item response with θ. In order to estimate $\rho'_{\theta i}$ from observable quantities, we can use the fact that it is also the loading of item i on the common factor of the tetrachoric item intercorrelation coefficients ρ'_{ij}. We note that a_i is a monotonic increasing function of $\rho'_{\theta i}$. Also that

$$\rho'_{\theta i} = \frac{a_i}{\sqrt{1 + a_i^2}} , \tag{7}$$

$$\rho'_{ij} = \frac{a_i}{\sqrt{1 + a_i^2}} \quad \frac{a_j}{\sqrt{1 + a_j^2}} . \tag{8}$$

Under the same assumptions, b_i can be estimated by using its relation to π_i, the proportion of correct answers to item i (Lord & Novick, 1968, section 16.9):

$$b_i = - \frac{1}{a_i} \sqrt{1 + a_i^2}\, \Phi^{-1}(\pi_i) , \tag{9}$$

where Φ^{-1} is the inverse of the normal ogive function defined in (3).

To keep matters simple in this preliminary survey, we will assume throughout the remainder of this report that all items available for a particular test or testing have the same value of a_i, and that all have the same value of c_i. Thus items will be chosen for administration solely according to their difficulty, as represented by the parameters b_i. It will be found that this simplification does not by any means make our problem a trivial one.

Before proceeding, let us write down a formula that will help us to interpret a_i in terms of more familiar test-theory statistics. Suppose

all items are <u>equivalent</u>, i.e., all items have the same icc; then all inter-item tetrachoric correlations ρ'_{ij} are the same. If also all $b_i = 0$ and all $c_i = 0$, then, assuming a normal distribution of θ and a normal ogive icc, the interitem phi coefficient ρ_{ij} (product moment correlation between dichotomously scored items) can be expressed as follows (Lord & Novick, 1968, eq. 15.9.3):

$$\rho_{ij} = \frac{2}{\pi} \arcsin \rho'_{ij} , \tag{10}$$

the angle being expressed in radians. By (8), since the items are equivalent,

$$\rho_{ij} = \frac{2}{\pi} \arcsin \frac{a_i^2}{1 + a_i^2} . \tag{11}$$

By the Spearman-Brown formula, the reliability of the number-right score on a test composed of n equivalent items is

$$\rho = \frac{n\rho_{ij}}{1 + (n - 1)\rho_{ij}} . \tag{12}$$

If $a_i = .333$, under the assumptions already made this reliability for a 60-item test will be .80; if $a_i = .5$, this reliability will be .90; if $a_i = 1.0$, this reliability will be .97. In view of this, we will choose $a_i = .5$ as a typical value and will address most of our attention to it.

The following table will help the reader to reinterpret the meaning of b_i. If θ is normally distributed with a mean of 0 and a standard deviation of 1, then, under the normal ogive model, the proportion of correct answers given by the group of examinees to an item with $a_i = .5$ is shown in the following table:

b_i =	-3.0	-2.0	-1.0	0	1.0	2.0	3.0
c_i = 0	.91	.81	.67	.50	.33	.19	.09
c_i = .2	.93	.85	.74	.60	.46	.35	.27

5. Stochastic Processes and Random Walks

As foreshadowed earlier, all the testings to be considered in the present report proceed one item at a time, as follows. After administration of the first item, each subsequent item is picked for administration by some predetermined rule, solely on the basis of the examinee's response to the preceding item. The choice among items is made entirely in terms of item difficulty, b. Let the superscript $v = 1,2,3,\ldots$ refer to the order in which the items are administered, so that item $v+1$ is the item administered immediately after item v.

Now, the origin and unit of measurement in which b and θ are expressed is purely arbitrary-- it is easily seen, for example, that adding a constant to b in (1), (2), (3), or (5) while subtracting the same constant from θ will have no effect on the item characteristic curve. Since we are free to choose an origin, we shall place it at $b^{(1)}$ so that hereafter

$$b^{(1)} \equiv 0 , \tag{13}$$

unless specifically stated otherwise.

In general, after a successful response we will want $b^{(v+1)} \geq b^{(v)}$; after an unsuccessful response, $b^{(v+1)} \leq b^{(v)}$. (Conceivably, blocks of items may be substituted for single items in this scheme, with some elaboration of the rule for choosing each successive block.)

Clearly, after the first item, the value of $b^{(v+1)}$, $v = 1,2,3,\ldots$, is a chance variable. The frequency distribution of $b^{(v+1)}$ depends, in accordance with (1), (2), (3), or (5), on the value of $b^{(v)}$ and on the examinee's value of θ (considered as fixed). Such a sequence of random variables is called a stochastic process. Furthermore, this process has the Markov property: Once $b^{(v)}$ is known for a given θ, the probability of any value of $b^{(v+1)}$ is independent of the values of $b^{(1)}, b^{(2)}, \ldots, b^{(v-1)}$. Thus for a given examinee the random variable $b^{(v)}$ constitutes a Markov process.

By what rule should the successive values of $b^{(v)}$ be chosen? A plausible branching rule would be the following. After administering item v, compute the maximum likelihood estimate of θ. Choose $b^{(v+1)}$ equal to this estimate. Do this for $v = 1,2,3,\ldots$.

For a fixed set of items, it is not difficult (at least not for a computer) to obtain a maximum likelihood estimate of θ from the likelihood function

$$L(u_1, u_2, \ldots, u_n) = \prod_{v=1}^{n} P_v(\theta)^{u_v} Q_v(\theta)^{1-u_v}, \tag{14}$$

where $P_v(\theta)$ is given by (1), (2), (3), or (5) and $Q_v(\theta) = 1 - P_v(\theta)$. Now for a fixed set of items, the values of the item difficulties are fixed and known; but for any stochastic process, they are random variables. This complicates the problem to such a point that we shall not attempt to evaluate the results obtained when the stochastic process itself depends on successive maximum likelihood estimation.

At this point, let us consider just one very simple kind of Markov process. This process is well known in bioassay work as the up-and-down

method. In our terms, if the examinee answers item v correctly, then we choose $b^{(v+1)} = b^{(v)} + d$, where d is some step size that we pick in advance; if he answers item v incorrectly, then $b^{(v+1)} = b^{(v)} - d$. It is apparent that in the up-and-down method, the random variable $b^{(v)}$ can take on the values

$$b^{(v)} = jd \tag{15}$$

where j is a (possibly negative) integer. Actually, j equals the number of correct responses minus the number of incorrect responses for the first $v - 1$ items. We see that

$$b^{(v+1)} = \begin{cases} b^{(v)} + d & \text{with probability } P(\theta - b^{(v)}), \\ b^{(v)} - d & \text{with probability } Q(\theta - b^{(v)}), \\ \text{any other value} & \text{with probability } 0, \end{cases} \tag{16}$$

where the notation $P(\theta - b^{(v)}) = P_v(\theta)$ is used in order to display the role of the item parameter $b^{(v)}$.

A Markov process in which the random variable can take only a denumerable set of values (as is the case here, see (15)) is called a Markov chain. A Markov chain satisfying (16), for specified values of P and Q is a random walk. The P's and Q's are called transition probabilities. The fact that for fixed $\theta - b^{(v)}$, $P(\theta - b^{(v)})$ does not depend on v is customarily described by saying that the transition probabilities are stationary.

Now,

$$b^{(v)} = b^{(1)} + d \sum_{r=1}^{v-1} (2u_r - 1) .$$

Let us note in passing that the likelihood function of the item responses u_g under the up-and-down method is

$$f(u_1, u_2, \ldots, u_n) =$$

$$\prod_{v=1}^{n} P^{u_v}[\theta - b^{(1)} - d \sum_{r=1}^{v-1} (2u_r - 1)]Q^{1-u_v}[\theta - b^{(1)} d \sum_{r=1}^{v-1} (2u_r - 1)] .$$

6. Scoring Methods

To set up a tailored testing operation, we must, in effect, choose not only a stochastic process but also a scoring procedure. For the most part, we shall consider just three simple possibilities, all of which have been used in experimental work on tailored testing. Any one of the different scores will be denoted by x. The total number of items to be administered to a given examinee is denoted by n; this is assumed fixed in advance of the testing. Let $u_i = 1$ denote a "correct" response to item i, and let $u_i = 0$ denote an incorrect response.

1. This score is the number of items answered correctly:

$$x = \sum_{v=1}^{n} u_v . \tag{17}$$

This is the conventional number-right score.

2. This score is the difficulty of the item that would have been administered to the examinee after the n-th item:

$$x = b^{(n+1)} = \begin{cases} b^{(n)} + d & \text{if } u_n = 1 , \\ b^{(n)} - d & \text{if } u_n = 0 . \end{cases} \tag{18}$$

It will be referred to as the final difficulty score.

3. This score is the average of the item difficulties, excluding $b^{(1)} = 0$ but including $b^{(n+1)}$ (as defined by eq. 18):

$$x = \frac{1}{n} \sum_{v=2}^{n+1} b^{(v)} . \qquad (19)$$

It will be called the <u>average difficulty score</u>.

To start with, let us consider the situation where we use the simplest of these scores, the final difficulty score, in conjunction with the up-and-down method of selecting items.

7. Up-and-Down Method with Final Difficulty Score

Since $b^{(v)}$ forms a Markov chain under the up-and-down method, standard formulas are available for finding the sampling distribution of the score $x = b^{(n+1)}$ for any specified n. These are outlined in the Appendix. For the present purpose of evaluating this particular tailored testing procedure, we will not need the entire frequency distribution of $x = b^{(n+1)}$, but we will need its expectation and sampling variance.

We need to improve our notation at this point. Let us write $X_v(b,\theta)$ to denote the score after administering v items when the difficulty of the first item administered is b and when the ability level of the examinee is θ. The score $X_v(b,\theta)$ is a random variable whose distribution depends on v, b, and θ. The distribution also depends on d and on the item parameters a and c, although these are not explicit in our notation. In actual practice, the first item administered has $b = 0$, as already explained in section 5. For the derivation, however, we will need to consider different possible values of b.

If the first item administered is answered correctly, the second item in the up-and-down method is picked to have a difficulty of $b + d$.

By virtue of the Markov property, once $X_1(b,\theta)$ is fixed, subsequent performance does not depend on the examinee's response to item 1. Thus when the first item is answered correctly, $X_{v+1}(b,\theta) = X_v(b + d,\theta)$ for $v = 1,2,3,\ldots$.

A similar analysis can be made for the case where the first item is answered incorrectly. Thus we can write for $v = 1,2,3,\ldots$

$$X_{v+1}(b,\theta) = \begin{cases} X_v(b + d,\theta) & \text{with probability } P(\theta - b) , \\ X_v(b - d,\theta) & \text{with probability } Q(\theta - b) . \end{cases} \tag{20}$$

This equation and similar ones derived by the same line of reasoning are fundamental to most of our practical results. It provides a relationship connecting the random variable X_{v+1} with the random variable X_v , allowing us to compute necessary quantities for X_n recursively.

Let

$$G_v(b,\theta) = X_v(b,\theta) - \theta \tag{21}$$

denote the error in the score X_v , and write $t = \theta - b$. Then, from (20) for $v = 1,2,3,\ldots$

$$G_{v+1}(b,\theta) = \begin{cases} G_v(b + d,\theta) & \text{with probability } P_t , \\ G_v(b - d,\theta) & \text{with probability } Q_t , \end{cases} \tag{22}$$

where we write P_t instead of $P(\theta - b)$. The values of P and Q are, as always, to be computed from (1), (2), (3), or (5) . In particular, we

see from (20) that

$$G_1(b,\theta) = \begin{cases} d - t & \text{with probability } P_t \ , \\ -d - t & \text{with probability } Q_t \ . \end{cases} \tag{23}$$

The bias of X_v is the expectation of G_v for given v and θ, which we denote by

$$E_v(b) = \mathcal{E}G_v(b,\theta) \ . \tag{24}$$

Although $E_v(b)$ is a function of θ, we will omit the symbol θ to keep the formulas simple. From (22), for $v = 1,2,3,\ldots$,

$$E_{v+1}(b) = P_t E_v(b + d) + Q_t E_v(b - d) \ . \tag{25}$$

From (23),

$$E_1(b) = (d - t)P_t - (d + t)Q_t = d(P_t - Q_t) - t(P_t + Q_t)$$

$$= d(1 - 2Q_t) - t \ . \tag{26}$$

The bias of X_n for given n and θ can be computed recursively using (26) and (25).

To get the sampling variance of $X_n = X_n(b,\theta)$, we start by defining W_v as the expected mean square error of X_v for a given θ :

$$W_v == \mathcal{E}G_v^2 = \mathcal{E}(X_v - \theta)^2 , \tag{27}$$

where for convenience we omit the arguments in parentheses. Thus, by (22),

$$W_{v+1}(b) = P_t W_v(b + d) + Q_t W_v(b - d) . \tag{28}$$

Also, by (23),

$$W_1(b) = (d - t)^2 P_t + (d + t)^2 Q_t . \tag{29}$$

The sampling variance of X_n is given by

$$\sigma^2_{X_n|\theta} == \mathcal{E}X_n^2 - (\mathcal{E}X_n)^2 = \mathcal{E}G_n^2 - (\mathcal{E}G_n)^2 = W_n(b) - [E_n(b)]^2 . \tag{30}$$

For future use, let us write down here one more recursion relation, enabling us to find

$$D_v(b) == \frac{\partial}{\partial\theta}\, \mathcal{E}X_v(b,\theta) . \tag{31}$$

From (20),

$$\mathcal{E}X_{v+1}(b,\theta) = P_t \mathcal{E}X_v(b + d,\theta) + Q_t \mathcal{E}X_v(b - d,\theta) ,$$

so that

$$D_{v+1}(b) = P_t D_v(b + d) + Q_t D_v(b - d) + [E_v(b - d) - E_v(b + d)]\frac{\partial Q_t}{\partial\theta} . \tag{32}$$

Also, from (20) and (13)

$$\mathcal{E}X_1(b,\theta) = P_t(b + d) + Q_t(b - d)$$

so that

$$D_1(b) = - 2d \frac{\partial Q_t}{\partial \theta} \quad . \tag{33}$$

If the icc is a normal ogive as in (3), then

$$\frac{\partial Q_t}{\partial \theta} = - a_i \phi[a_i(\theta - b_i)] \quad .$$

8. Evaluation of Estimates of θ

Before giving numerical results evaluating the up-and-down method with final difficulty score, it is necessary to consider at some length just how such results should be evaluated. This is not a trivial matter.

Empirical research studies (see Linn, Rock, & Cleary, 1969, and other references given there; also Hansen & Schwarz, 1968) have often used the correlation of tailored-test score x with some outside criterion to evaluate the effectiveness of testing and scoring procedures. If a particular examiner has repeatedly tested similar groups, he may know approximately in advance the distribution of θ in the next group he plans to test; such an examiner may well use group statistics and Bayesian methods. Although these kinds of evaluation have obvious face validity, they will not be used here for at least two reasons:

1. The correlation coefficient is a group statistic, whereas the problem here is to determine the accuracy with which we can measure a single individual. Important information about the accuracy obtainable for specific individuals is lost when this information is pooled over individuals to get a group statistic.

2. In order to obtain a group statistic, we would have to make some assumption about the frequency distribution of θ in the group studied. Such an assumption would prevent easy generalization of our results to groups with substantially different distributions of θ. (The few results available to date (see Lord, 1968b) run against the convenient assumption that θ is likely to be approximately symmetrically distributed at least in the case of highly selected groups such as college students.)

The gain to be hoped for from tailored tests arises entirely (or nearly so) from tailoring the item difficulties to the ability of the examinee. But in a typical test that is not too heterogeneous in item difficulty, most of the items are already well tailored to the abilities of most of the examinees. Thus tailored testing can not provide greatly improved measurement for most examinees. The value of tailored tests is primarily for those examinees for whom the conventional test would be too easy or too difficult. The correlation coefficient over the entire group of examinees is not a good index for judging the improved measurement gained by a minority.

One way to describe the measurement properties of the score x is by giving its standard error (the square root of the sampling variance $\sigma^2_{x|\theta}$ of eq. 30). It is no surprise to find that the standard error depends on the unknown value of θ. The measurement properties of the score x cannot be summarized by a single number, but must be represented by a curve--a function of θ. Thus we might find, for example, that score x_1 provides

more accurate measurement than score x_2 for a certain range of values of θ, while score x_2 is more accurate for examinees outside this range.

In bioassay work (e.g., Brownlee, Hodges, & Rosenblatt, 1953), $E_n(b)$, the bias of x, must be taken into account. This is done by using the expected mean-square error $W_n = \mathcal{E}(X_n - \theta)^2$ to describe the accuracy of measurement. As shown in (30), the expected mean-square error exceeds the sampling variance by the square of the bias.

In mental testing, on the other hand, the scale in which θ is measured has an arbitrary origin and unit of measurement. Thus a constant bias in the score x, or even a bias that changes linearly with θ, would not impair the value of x at all. To carry matters further, the scale for θ that yields (2) or (5) is highly arbitrary. Any monotonic transformation of this scale would be defensible for measuring the examinee. In comparing different tailored testing procedures, we must use an index of effectiveness that always leads to the same conclusion no matter what monotonic transformation of the θ scale is chosen.

In order to describe the effectiveness of the score x for measurement purposes, we will use

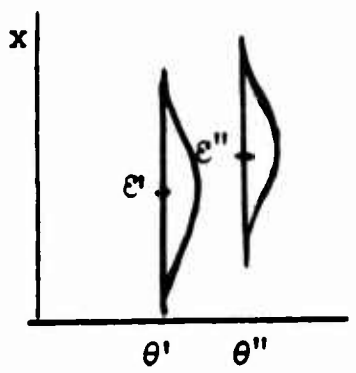
$$I_x(\theta) = \left(\frac{\frac{\partial}{\partial\theta}\mathcal{E}(x|\theta)}{\sigma_{x|\theta}} \right)^2 = \frac{D_n^2(b)}{\sigma^2_{X_n|\theta}} . \qquad (34)$$

(The reader may wish at this point to glance at Figure 2, which shows $\sqrt{I_x(\theta)}$ for certain testing procedures.) This is the quantity recommended by Birnbaum (1968, eq. 17.7.10) to measure the "information" in the score x. The use of $I_x(\theta)$ also was recommended by Lord (1952, eq. 57) and, in a very different context, by Mandel and Stiehler (1954). We will call $I_x(\theta)$ the information function for the score x.

As Birnbaum shows, in large samples, $I_x(\theta)$ is inversely proportional to the square of the length of the confidence interval for estimating θ from x . Birnbaum uses this information function for small as well as large samples, and we shall do so here. The meaning and justification of this index in small samples are well described by Mandel, whom we paraphrase closely here:

> If it is desired to differentiate between two nearby values, θ' and θ'' , by means of the corresponding measurements x' and x'' , it is apparent that the success of the operation will depend on two circumstances: (1) the magnitude of the difference $\varepsilon'' - \varepsilon' = \varepsilon(x''|\theta'') - \varepsilon(x'|\theta')$ for a given difference $\theta''- \theta'$, i.e., the magnitude of the slope $(\varepsilon'' - \varepsilon')|(\theta'' - \theta')$; and (2) the precision of measurement $\sigma_{x|\theta}$. These two desiderata can be combined in a single criterion, the information function, defined as the square of the ratio of the slope to $\sigma_{x|\theta}$.

It is helpful to visualize the situation with the aid of a small diagram:



A more formal discussion of the small-sample interpretation is given by Lord (1952).

It is important to note that $I_x(\ )$ is an operator, not a function. This means that $I_x(a\theta)$ must be found from the definition in (34), <u>not</u> by writing down $I_x(\theta)$ and then substituting $a\theta$ for θ .

It is apparent from (34) that any change in the unit used to measure θ does change $I_x(\theta)$. Thus $I_x(\theta)$ is not a pure number. In fact, $1/\sqrt{I_x(\theta)}$ is expressed in the same score units as θ.

Suppose that a monotonic increasing transformation is made on θ so that $\theta^* == \theta^*(\theta)$ replaces θ. Then the denominator of (34) remains unchanged, but the numerator must be multiplied by $(\frac{\partial\theta}{\partial\theta^*})^2$ in order to find $I_x(\theta^*)$. Since $\frac{\partial\theta}{\partial\theta^*}$ may have any form, it is possible to make $I_x(\theta^*)$ assume any shape desired, within the restriction $I_x(\ ) > 0$, simply by a suitable choice of the transformation θ^*.

The conclusion to be drawn is that unless we are willing to assert that we have used a uniquely appropriate scale for measuring θ, we cannot draw conclusions from the shape of the information function. This drastic limitation leads to no difficulties in the present study since we shall always be comparing two or more information functions, all based on the same scale for θ. The ratio between the information function for x_1 and the information function for x_2 measures the <u>relative efficiency</u> of x_1 and x_2 for estimating θ. <u>The relative efficiency is unaffected by any differentiable monotonic increasing transformation of</u> θ, since the factor $\frac{\partial\theta}{\partial\theta^*}$ appears twice and cancels out.

In studying the effectiveness of tailored testing procedures, it will be helpful if we can compare them to more familiar procedures. To keep matters simple, we will compare the information function of the tailored testing procedure with the information function of the number-right score on a conventional test composed of n equivalent items with normal ogive icc and with all $b_i = 0$. This test will hereafter be referred to as the <u>standard test</u> (use of number-right score is to be understood). The information function for (number right score on) this standard test is

(Birnbaum, 1968, eqs. 20.2.2, 20.5.1)

$$I(\theta) = \frac{nP'_\theta(\theta)^2}{P(\theta)Q(\theta)} \quad , \tag{35}$$

where $P'_\theta(\theta)$ is the derivative of $P_i(\theta)$ with respect to θ. The subscript x has been dropped on the left of (35) because for tests composed of equivalent items, the number-right score is a sufficient statistic for estimating θ (Birnbaum, 1968, section 18.3.1); consequently, (35) represents the maximum information that could be obtained from the responses to the n items by any scoring method.

Since these n items are ideally suited for an examinee at $\theta = 0$, it follows that no information curve can ever be higher at any value of θ than the value given by (35) when $\theta = 0$. <u>This provides a horizontal line at or below which all information curves must fall</u>. Note that this limit is a result of the assumption that a_i is the same for all items. It might well be found in practice that the harder items tend to have higher or lower a_i than the easier items, in which case the limit would no longer be a horizontal straight line. The limiting curve can still be computed from (35) if the a_i are known.

It is worth mentioning that the information function for the number-right score on any conventional test is proportional to n, the length of the test. Thus, <u>a percent increase in any information function, achieved by whatever means, can be understood as an increase in information equivalent to that obtained for a number-right score on a conventional test by increasing the number of test items by the same percentage</u>.

9. Information Functions for the Up-and-Down Method with Final Difficulty Score

Even when we restrict attention just to the up-and-down method with final difficulty score, there is still quite an assortment of possibilities to be investigated. Here, we restrict our attention to items with normal ogive characteristic curves (for some results under the logistic model, see section 13). In the next several sections, we will consider only the case where $c = 0$.

This still leaves us with four parameters to take into consideration: a , n , d , and θ . Figure 2 shows information functions for the up-and-down method with final difficulty score, computed from (34), (32), and (30), for $n = 10$ and for $n = 60$. Figure 2 is appropriate for any value of a_i in a wide range. This is possible because dividing a_i by a constant does not change the value of the icc in (3) (or in (1), (2), or (5), for that matter) provided b_i and θ are multiplied by the same constant.

If $a_i = a = 1$, we are likely to be interested in the range of θ between $\theta = -3/a = -3$ and $\theta = 3/a = +3$, say. Since we have set $\sigma_\theta = 1$ (see section 4), we may expect that not too many people will lie outside some range such as $\pm 3\sigma$. That part of Figure 2 beyond $a\theta = \pm 3$ is shown here for completeness; however, it has been shaded to indicate that it is probably only rarely of interest.

If we consider items with $a_i = a = .50$ instead of 1.0, then, if other things are the same as above, since $\sigma_\theta = 1$ always, we shall be interested in the range from $\theta = -1.5/a = -3$ to $\theta = +1.5/a = +3$. When we halve the value of a_i , we must double the value of b_i (as well as of θ) if P_i is to remain unchanged. This means that for $a_i = a = 1$, the

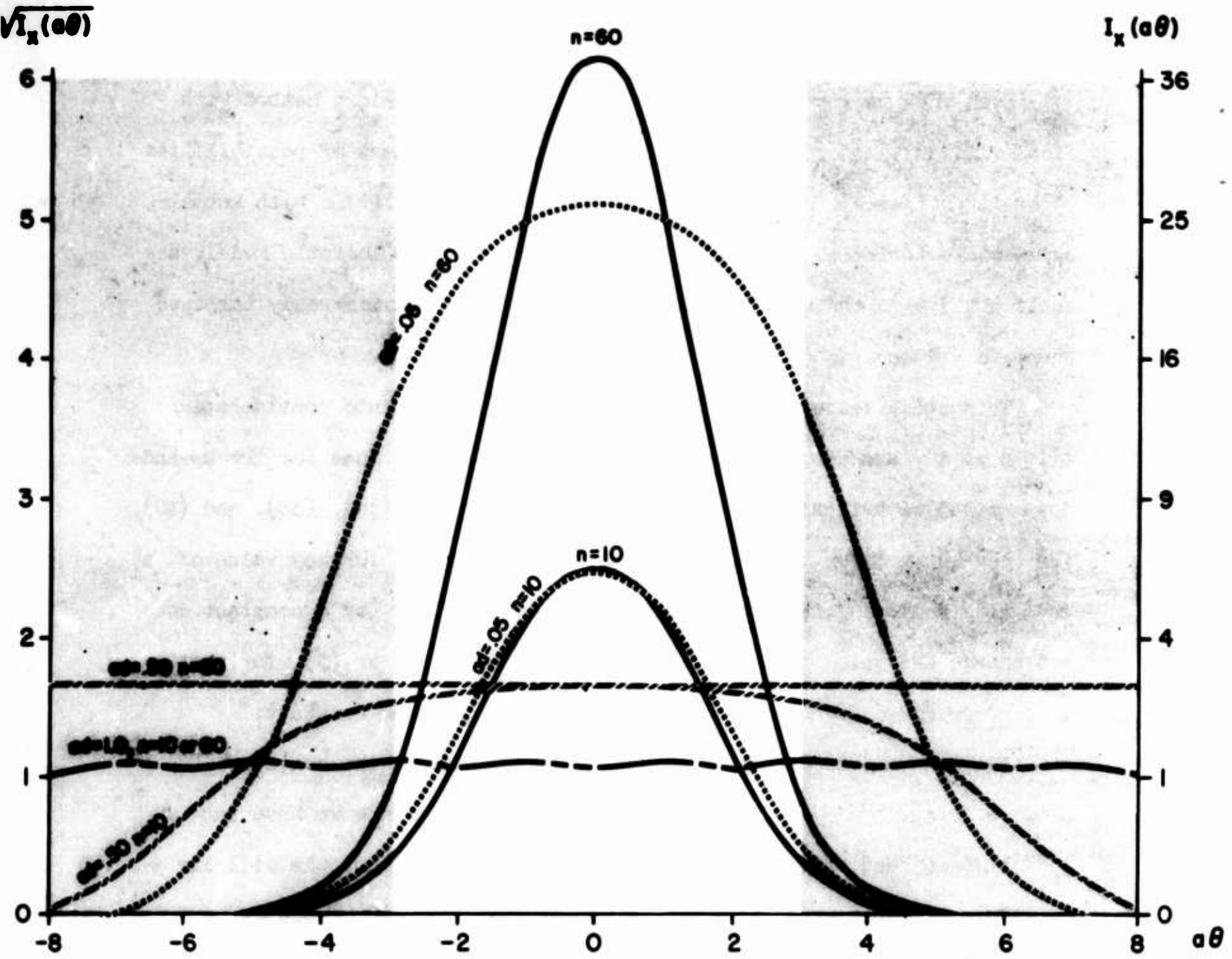


Fig. 2. Information functions for the up-and-down method with final difficulty score (solid lines are for the standard tests).

curve labeled $ad = 1.0$ represents a random walk procedure with step size $d = 1.0/a = 1.0$; but when $a_1 = a = 0.5$, the step size is now $d = 1/a = 2.0$.

Vertical distances in Figure 2 and in all similar figures are proportional to the square root of the information function. Thus vertical distance is directly proportional to a large-sample approximation for the length of the confidence interval for estimating θ from test score. The vertical scale on the right, however, is numbered to show the information function, not its square root. Thus the numbers given on the scale on the right are proportional to the number of test items required to produce a corresponding amount of information by conventional testing methods.

The two solid lines labeled $n = 10$ and $n = 60$ represent the information functions of a 10-item and a 60-item standard test (see section 8), as given by (35). If we read the heights of these curves on the right-hand scale, we can confirm that the 60-item standard test gives exactly 6 times as much information as the 10-item standard test, regardless of the value of θ .

The numerical results shown in Figure 2 and in subsequent figures were obtained by programming a computer to evaluate equations (25), (28), and (32) recursively for $v = 1,2,\ldots,59$. All results obtained recursively, both here and in later sections, were independently checked up through $n = 10$ (thus checking the formulas also) by an entirely separate computer program that computed the probability of each of the 2^{10} possible patterns of response to $n = 10$ items, computed the score for each pattern, and then computed the mean and the variance of these scores over all possible patterns, and also the derivative $D_n(b)$. This latter method of computing results cannot be extended much beyond $n = 10$ or 12 .

The details of interpreting Figure 2 are considered at some length here because most subsequent numerical results will be presented similarly.

Hereafter, the verbal interpretation will be for the most part limited to the case where $a_i = a = .50$, in order to keep the presentation from getting out of hand.

The scalloped effect when the step size is 2.0 (when $ad = 1.0$) is due to the fact that when n is even, the difficulty of the $(n + 1)$-th item must be an odd multiple of 2.0. Thus examinees at $\theta = \pm 2.0$ or at $\theta = \pm 6.0$ are measured more accurately than those at intermediate levels. Actually, all curves shown here and in subsequent figures should be scalloped. However, the effect is hardly noticeable for smaller values of d and has been ignored in drawing the figures.

When step size is $d = 2.0$, the amount of information obtained within the range $-8 \leq a\theta \leq +8$ is not appreciably increased by additional testing after the first 10 items-- the same curve adequately represents the information functions for both $n = 10$ and $n = 60$. Clearly the step size is too large to provide accurate measurement. The same is true for $d = 1.0$ $(ad = 0.5)$ within the range $-3 \leq \theta \leq +3$ but not outside this range.

When step size is reduced to $d = 0.10$ $(ad = 0.05)$, the information obtained near $\theta = 0$ is greatly increased, less so for θ at ± 3.0 $(a\theta = \pm 1.5)$.

For $a = 0.5$, the 10-item tailored testing procedure with $d = .10$ is better than the standard test for almost all θ. For $n = 60$, the standard test is better at $\theta = 0$, but its effectiveness falls off, so that at $\theta = \pm 3.0$ $(a\theta = 1.5)$, it provides only about 80 percent as much information as does the tailored test with $d = 0.10$. We see here the broad outlines of a basic problem in tailored testing. We need a small step size to compete with the accuracy of measurement provided by the standard test for typical individuals (near $\theta = 0$), but we need a large step size in order to obtain accurate measurement

of atypical individuals (at $\theta = \pm 3$, say). The optimal step size in any situation of course depends on a_i, on n, and on the accuracy of measurement required at different levels of θ.

Note that no information curve can ever be higher at any value of θ than the maximum of the information curve for the standard test.

10. The Up-and-Down Method with Number-Right Score

Keeping the same random-walk method of sequencing items, let us see to what extent we get effective measurement when we change the examinee's score from $b^{(n+1)}$, the difficulty of the $(n+1)$-th item, to $\sum_{v}^{n} u_v$, the total number of items answered correctly. This number-right score has been used in experimental studies of tailored testing. Denoted by X_n or by x, it is the score referred to throughout this section, unless otherwise specified. The reader may wish before reading further to form his own judgment as to the relative effectiveness of the scores $b^{(n+1)}$ and $\sum_{v}^{n} u_v$.

Let $X_v(b,\theta)$ now denote the score $\sum_{r=1}^{v} u_r$ obtained in the up-and-down method when the difficulty of the first item administered is b and when the ability of the examinee is θ. As in section 7, we have a recursion equation for the random variable X:

$$X_{v+1}(b,\theta) = \begin{cases} X_v(b+d,\theta) + 1 & \text{with probability } P(\theta - b), \\ X_v(b-d,\theta) & \text{with probability } Q(\theta - b), \end{cases} \tag{36}$$

for $v = 1,2,3,\ldots$. From this we can derive equations for $\mathcal{E}X_{v+1}(b,\theta)$, for $\mathcal{E}X^2_{v+1}(b,\theta)$, and for $D_{v+1}(b)$ similar to (25), (28), and (32). The

error variance $\sigma^2_{X_n|\theta}$ can be computed from the first line of (30). The information function can be computed from (34).

We find that the bias, the mean squared error, and the sampling variance of $\Sigma^n u_v$ are (not surprisingly) very different from the corresponding quantities for $b^{(n+1)}$. But we find that the information functions for the two scoring methods are exactly the same!

This result leads us to examine more closely the relation between $b^{(n+1)}$ and $\Sigma^n u_v$. If an examinee starts at $b^{(1)} = 0$ and finishes at $b^{(n+1)}$ after n steps each of length d, it is clear that for $d > 0$ his number of right answers exceeds his number of wrong answers by $b^{(n+1)}/d$. Since the number of right answers plus the number of wrong answers equals n, we have for $d > 0$

$$\frac{b^{(n+1)}}{d} = \sum_{v=1}^{n} u_v - (n - \sum_{v=1}^{n} u_v)$$

or

$$b^{(n+1)} = d(2 \sum_{v=1}^{n} u_v - n) . \tag{37}$$

<u>In the up-and-down method there is a linear relationship between final difficulty score and number-right score</u>. Thus although the scores show very different biases and sampling variances, they are equally effective for measurement purposes.

11. Bioassay

In bioassay work, the average difficulty score $\frac{1}{n} \sum_{v=2}^{n+1} b^{(v)}$ is commonly recommended for the up-and-down method. A considerable amount of theoretical work has been done for this method: Tsutakawa (1967a, 1967b,

1963) was primarily concerned with asymptotic results. Wetherill (1963) used Monte Carlo methods to investigate a wide variety of branching and scoring methods empirically. Dixon (1965), Cochran and Davis (1964), Brownlee, Hodges and Rosenblatt (1953), Dixon and Mood (1948), and others have derived useful large- and small-sample formulas. They also obtained numerical results, mostly for $n = 10$ or 12.

The method of approach and most of the equations of section 12 are simple extensions of those in the cited work of Brownlee, Hodges, and Rosenblatt. This same general approach is used in several other sections, including section 7.

It will be worthwhile to point out some of the similarities and differences between our problem and the bioassay problem, as commonly formulated. The bioassayist typically starts with equation (1) or (3), just as we have here. Whereas we control a_i and b_i while trying to estimate θ, the bioassayist controls θ while trying to estimate the value of b and (sometimes) the value of a. For him θ might be the dosage of the insecticide applied, for example, in which case b would be the LD50, the dosage at which 50 percent of the treated insects die. The bioassayist chooses the dose $\theta^{(1)}$, administers it to one insect (or to several), and observes the response: survival ($u_g = 0$) or death ($u_g = 1$). He then chooses a dose, $\theta^{(2)}$, administers it to another insect, and continues in this way.

Whereas we are usually only interested in the <u>relative</u> values of θ for different examinees, and often only in the rank order of these values, the bioassayist must estimate the absolute value of b for a single given insecticide. Thus the bioassayist uses the mean squared error as a criterion of effective estimation, whereas we use the information function.

Bias is a serious problem for the bioassayist, whereas it is usually of no concern to us.

The fact that the bioassayist has two unknown parameters, a and b, creates a very serious problem. He must choose a step size d without knowing a. If he picks d too large, the sampling error of his estimate of b will be excessive, even for sizable n. On the other hand, if d is small and a happens to be small also, the true value of b may be so far from the value $\theta^{(1)}$ at which the bioassay is started that $\theta^{(n)}$ can never reach the value b in n steps of length d. This results in an unacceptable bias in the estimate of b. Without some knowledge of a, there is no entirely safe way of choosing d. It is possible to estimate a from the observations themselves as they accumulate, but these estimates are very unreliable for the values of n frequently used in bioassay.

Most work on the up-and-down method assumes that a is known or else that a can be bounded (from previous experience) within certain limits. In the latter case, the step size must be chosen uncomfortably high, to allow for the possibility that a may be small. Here, we have assumed that in mental testing the necessary item parameters have all been determined by pretesting, with good accuracy. The result of all this is that we will be able to use a smaller step size in mental testing than is commonly recommended for the up-and-down method in bioassay. This, together with the fact that bias is usually of no concern to us, will allow us to obtain better results from the up-and-down method than are usually possible in bioassay.

12. The Up-and-Down Method with Average Difficulty Score

Dixon and Mood (1948), starting with the normal-ogive model for bioassay, derived approximations to the maximum likelihood estimators for a and b. Brownlee, Hodges, and Rosenblatt (1953) proposed using $\frac{1}{n}\sum_{v=2}^{n+1} \theta^{(v)}$ to estimate b in the bioassay problem, pointing out that this estimator is asymptotically equivalent to the one recommended by Dixon and Mood. Our average difficulty score $\frac{1}{n}\sum_{v=2}^{n+1} b^{(v)}$ corresponds directly to the estimator used by Brownlee, Hodges, and Rosenblatt. For the most part, we will study exact small-sample properties of the average difficulty score rather than asymptotic properties. The development given here for $H = L = 1$ is essentially the same as that of Brownlee, Hodges, and Rosenblatt, except that they are not concerned with $\sigma_{x|\theta}$, nor with quantities analogous to $D_v(b)$, nor with the information function $I_x(\theta)$.

In this section, x will always refer to the average difficulty score. By definition, the random variable $X_v(b,\theta) = \sum_{r=2}^{v+1} b^{(r)}$ will be the sum of item difficulties obtained under the up-and-down method when the first item is of difficulty b and the examinee is at ability level θ. The basic recursion, corresponding to (20) for the final difficulty score, is seen to be

$$X_{v+1}(b,\theta) = \begin{cases} b + Hd + X_v(b + Hd, \theta) & \text{with probability } P_t, \\ b - Ld + X_v(b - Ld, \theta) & \text{with probability } Q_t, \end{cases} \tag{38}$$

where $H = L = 1$ (the symbols H and L are not needed here, but will be useful in later sections).

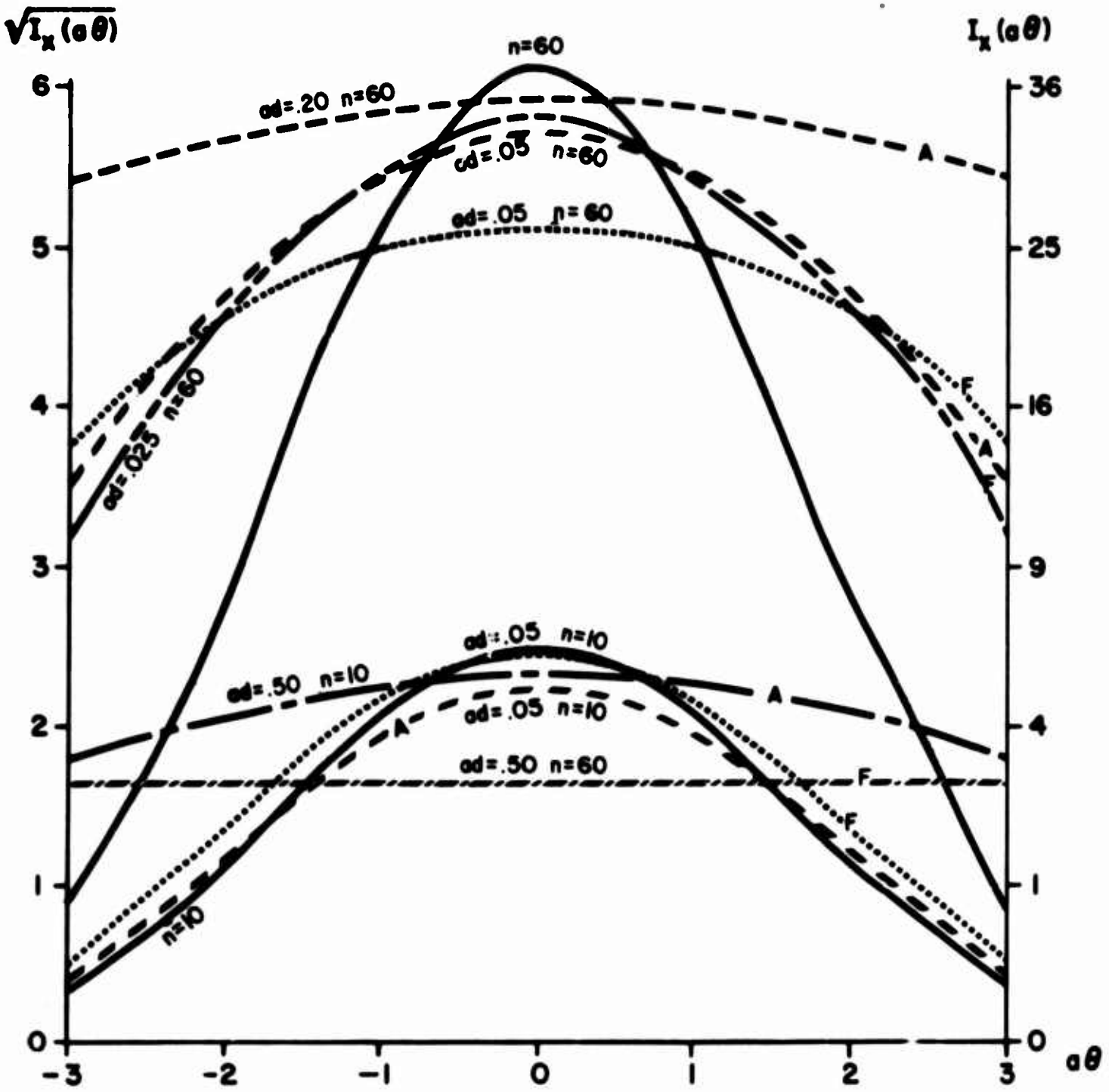


Fig. 3. Information functions for the up-and-down method, comparing final difficulty score (F) and average difficulty score (A).

Since X_v is a sum, not an average, let us define $G_v(b,\theta)$ by

$$G_v(b,\theta) == X_v(b,\theta) - v\theta \ .$$

Then by the same reasoning used in section 7

$$E_{v+1}(b) = P_t E_v(b + Hd) + Q_t E_v(b - Ld) + E_1(b) \ , \tag{39}$$

$$E_1(b) = (Hd - t)P_t - (Ld + t)Q_t \ . \tag{40}$$

Similarly, the mean squared error is found from

$$W_{v+1}(b) = P_t W_v(b + Hd) + Q_t W_v(b - Ld) + W_1(b)$$

$$+ 2P_t(Hd - t)E_v(b + Hd) - 2Q_t(Ld + t)E_v(b - Ld) \ , \tag{41}$$

$$W_1(b) = (Hd - t)^2 P_t + (Ld + t)^2 Q_t \ . \tag{42}$$

Finally,

$$D_{v+1}(b) = D_1(b) + P_t D_v(b + Hd) + Q_t D_v(b - Ld)$$

$$+ [E_v(b - Ld) - E_v(b + Hd)] \frac{\partial Q_t}{\partial \theta} \ . \tag{43}$$

$$D_1(b) = - \partial(H + L) \frac{\partial Q_t}{\partial \theta} \ . \tag{44}$$

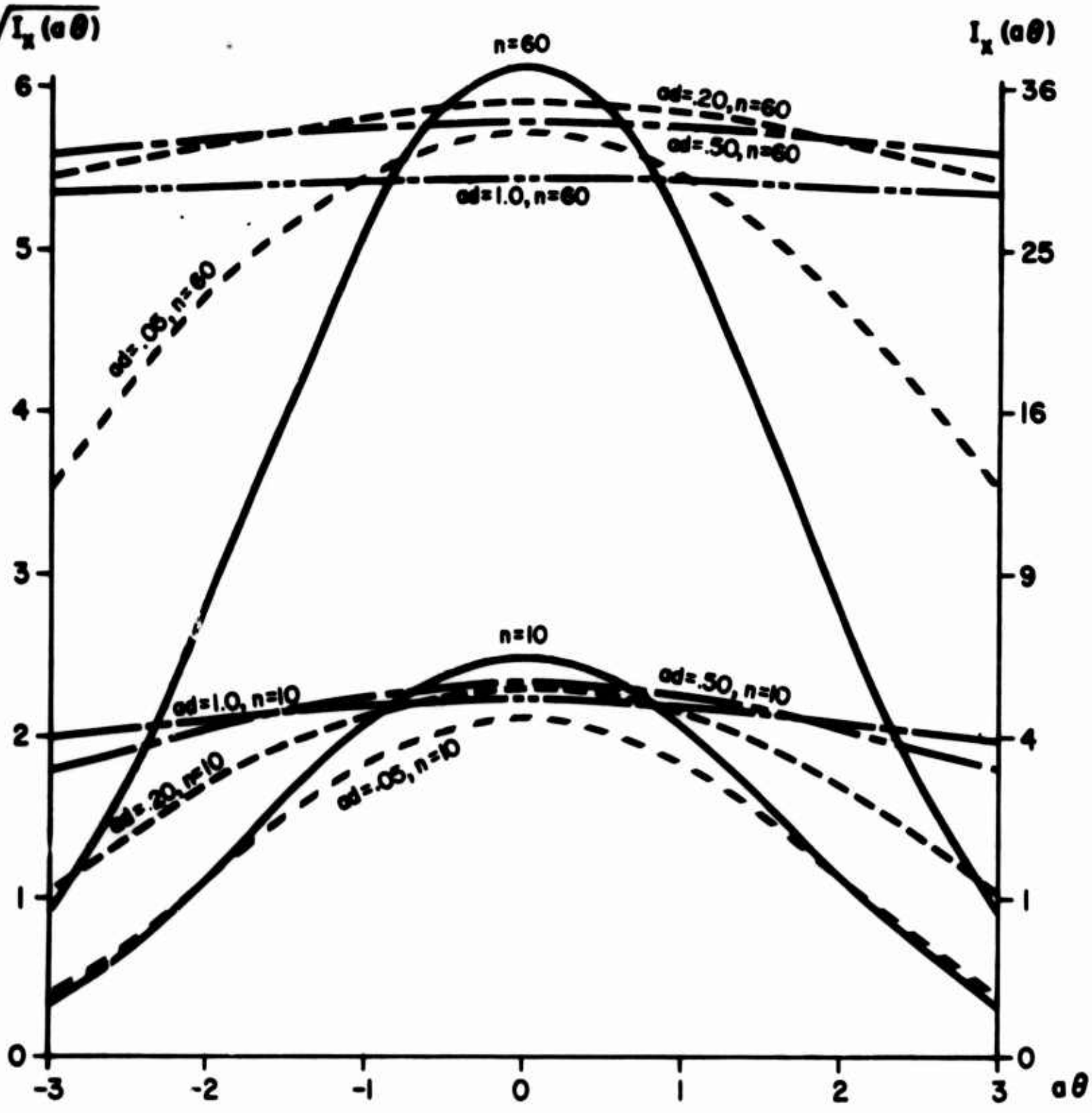


Fig. 4. Information functions for the up-and-down method with final difficulty score.

If the icc is a normal ogive with or without $c = 0$, as in (3) or (5), then

$$\frac{\partial Q_t}{\partial \theta} = -a_i(1 - c_i)\phi[a_i(\theta - b_i)] \ . \qquad (45)$$

Figure 3 compares information curves for some final difficulty scores (marked F) with those for the corresponding average difficulty scores (marked A) both when $n = 10$ and when $n = 60$. When $n = 10$, the final difficulty score with $ad = .05$ is best when $\theta = 0$; however, effectiveness has fallen off when $a\theta = \pm 1.5$. The average difficulty score with $n = 10$, $ad = .50$ is not quite as good when $\theta = 0$, but is distinctly better when $a\theta = \pm 1.5$.

When $n = 60$ and $ad = .05$, the average difficulty score is better than the final difficulty score throughout the range $-2 \leq a\theta \leq 2$. The final difficulty score can be made to provide more information near $\theta = 0$ by shortening the step size (the curve for $ad = .025$ is shown). However, as the step size is shortened, the information curve for the final difficulty score must approach the curve for the standard test, shown as a solid line. This is so because final difficulty score is perfectly correlated with number-right score (eq. 37). Thus, shortening the step size produces only small gains near $\theta = 0$ and ultimately leads to a serious loss of accuracy for final difficulty score at more extreme values of θ.

The average difficulty score with $ad = .20$ is better than any of the other tailored procedures throughout the entire range shown. It is almost as good as the standard test near $\theta = 0$ and is better for other values of θ. This result cannot be improved by shortening the step size. It will be seen that when the average difficulty score is used, too short a step size causes loss of accuracy throughout the entire range of θ.

We conclude tentatively that for the up-and-down method, the average difficulty score is preferable to the final difficulty score, at least whenever we want good measurement throughout a range of θ, not just at $\theta = 0$. There may be some exceptions to this in the case of very short tests. Our further investigations of the up-and-down method will be directed principally at 60-item tests and will be based entirely on average difficulty scores.

All the information curves in Figure 4 relate to the up-and-down method with average difficulty score, except for the solid curves, which show the information produced by the standard tests. When $n = 10$, the best step size seems to be between $d = .2/a$ and $d = .5/a$. With this step size the tailored test is 85 percent efficient at $\theta = 0$ compared to the standard test and is more efficient than the standard test for extreme values of θ. When $n = 60$, the best step size seems to be roughly $d = .2/a$. With this step size, the tailored test is more than 90 percent efficient at $\theta = 0$ compared to the standard test. At $a\theta = \pm 1.5$, the standard test is only 48 percent efficient compared to this tailored test (i.e., the standard test would have to be more than twice as long as the tailored test in order to produce the same amount of information for examinees at $\theta = \pm 1.5/a$).

It is no surprise that step size can be too small for effective measurement of examinees at extreme values of θ -- n cumulated steps may never reach the item difficulty level appropriate for the examinee. It may well seem surprising, however, that when average difficulty score is used, step size can be too small for measuring examinees at $\theta = 0$, as shown in Figure 4. After all, the most effective measurement for such examinees in conventional testing is obtained when $b_i = 0$ for all items.

Consideration of a limiting case of the up-and-down method, the case where the step size is zero, may throw some light on the apparent paradox. If the step size is zero, all examinees will have the same average difficulty score, regardless of their θ. In this limiting case, it is clear that number-right score can provide a good measure of θ whereas average difficulty score can not. This suggests that the effectiveness of average difficulty score at $\theta = 0$ falls off when step size becomes too small, as illustrated by Figure 4.

13. A Comparison of Logistic and Normal Ogive Item Characteristic Curves

Except for this section, all information functions given in this report were calculated under the normal-ogive model given by (3) or (5). Here, in Table 1, we compare information functions obtained from (3) for normal-ogive icc's with those obtained from (1) for logistic icc's.

The numerical differences between the two models are not wholly negligible. However, it does not seem necessary to compute all information functions under both normal-ogive and logistic models.

14. The Effect of Chance Success

Common sense, and also empirical data, tells us that even examinees at the lowest θ levels have a chance considerably greater than zero of getting correct answers to multiple-choice questions. Clearly, this will result from guessing, whether random or nonrandom (if there is such a thing as "nonrandom guessing").

The three-parameter models (2) and (5) were designed to fit this situation. The parameter c_i represents the probability of success for

Table 1

A Comparison of Information Functions for

Normal-Ogive and Logistic Icc's

Step Size $\underline{ad}$	Test Length $\underline{n}$	Icc	Information function $I_x(\theta)$ at $\theta = 0$	± 1	± 2	± 3
1.0	60	normal	5.45	5.45	5.40	5.37
		logistic	5.44	5.43	5.38	5.34
	10	normal	2.22	2.21	2.10	2.00
		logistic	2.23	2.20	2.07	1.97
0.05	60	normal	5.73	5.48	4.72	3.53
		logistic	6.10	5.60	4.41	2.97
	10	normal	2.27	1.97	1.20	0.45
		logistic	2.42	1.87	0.98	0.44

low-level examinees. There is no need to specify any particular relation between c_i and the number of possible responses to a multiple-choice item. With 5-choice items, practical experience indicates that most items will be fitted by values of c_i between .10 and .20. There is no need to assume that all items in a test have the same c_i. However, for simplicity, we here assume that all items have $c_i = .20$. We will investigate whether or not this has any clear-cut implications for tailored testing.

Figure 5 compares information functions for $c_i = 0$ with those for $c_i = .20$ for all items. As always, the standard tests are shown by solid lines. All other curves are for the up-and-down method with average difficulty score, $n = 60$.

The figure confirms that chance success in answering items seriously reduces measurement efficiency. The loss is, of course, greatest at low θ levels, but in tailored testing it is substantial at all θ levels. The loss in information at $\theta = 0$ is 33 percent for the standard test, 38 percent for $ad = .05$, 46 percent for $ad = .20$, and 68 percent for $ad = 1.0$. It appears that the larger the step size, the greater the loss due to guessing.

The simple up-and-down method as described here is designed to move towards a situation where about half the items administered to an examinee will be answered correctly, about half incorrectly. This is the proper ratio when there is no chance success, but it has long been recognized that easier items are preferable when chance success occurs. This conclusion is particularly obvious when $c_i \geq .5$, so that 50-percent success is at or below the chance level. It used to be thought that for optimum measurement the examinee should answer $(1 + c)/2$ of the items correctly. This would be 60 percent of the items, for our case where $c = .2$.

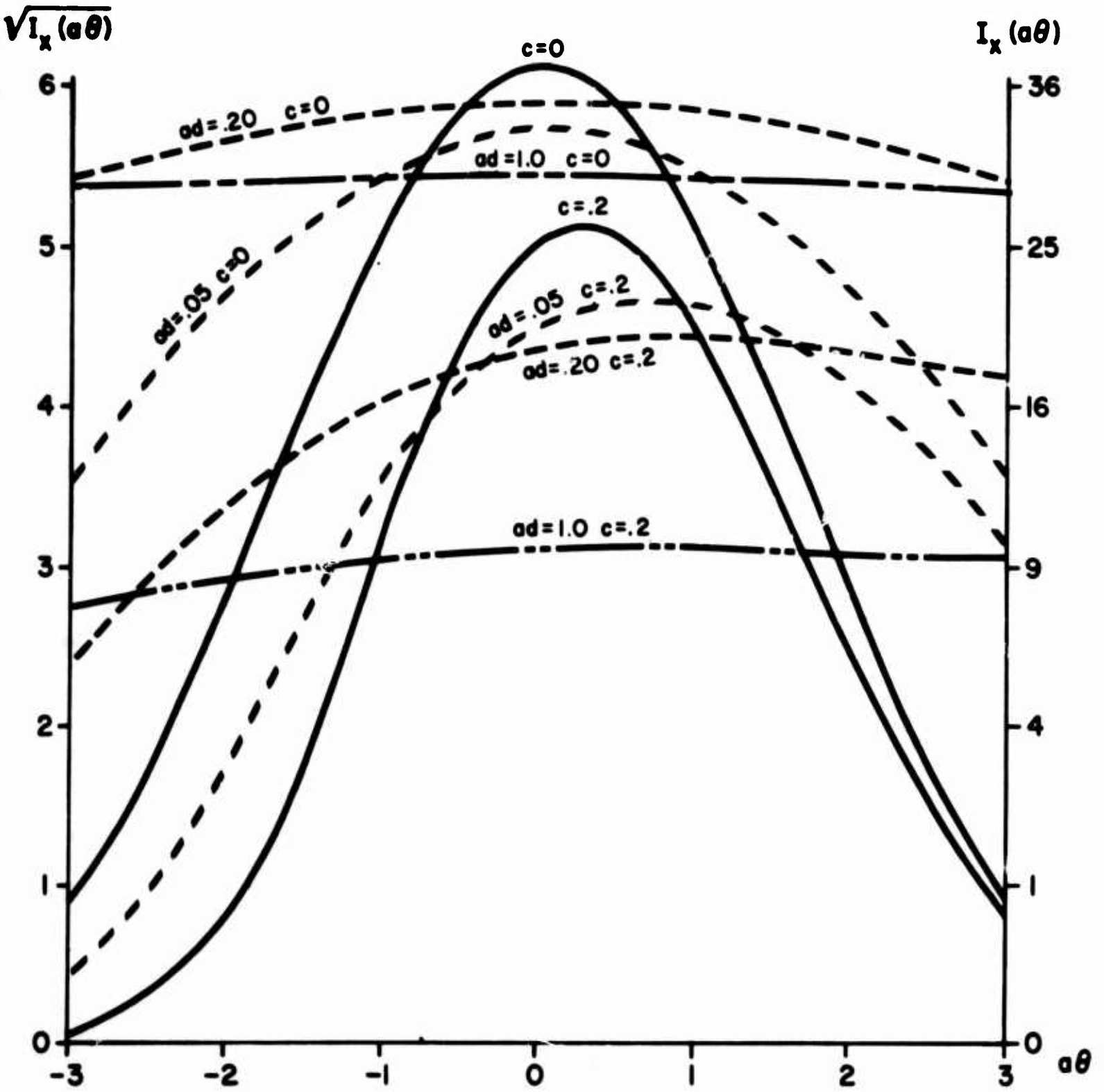


Fig. 5. A comparison of information functions when there is chance success (c = .2) and when there is not (c = 0).

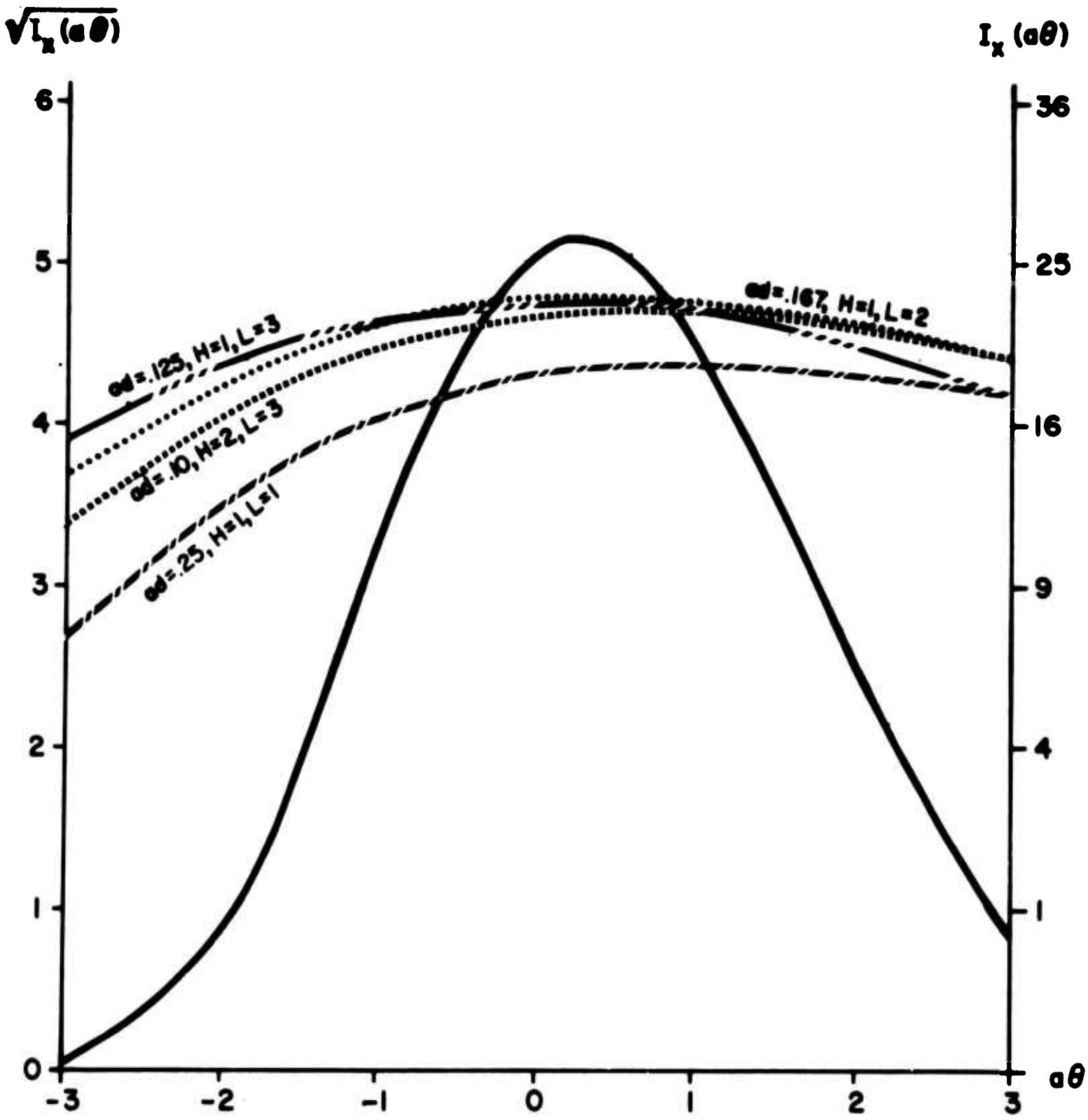


Fig. 6. A comparison of various H-L methods.

Lord (1953, pp. 67-69) found that still easier items than this would be preferable. A formula provided by Birnbaum (1968, eq. 20.4.22) for the logistic model shows that for optimum measurement we should have

$$b_i - \theta = -\frac{1}{1.7a_i} \log_e \frac{1 + \sqrt{1 + 8c_i}}{2} \quad . \tag{46}$$

When $a_i = .5$ and $c_i = .2$, we find that $b_i - \theta$ should be -.314, in which case, by (2), $P_i(\theta) = .653$.

How shall we arrange matters so that on a sufficiently long test the examinee will eventually be answering 60 to 70 percent of the items correctly? One possibility is to make the step size in the positive direction smaller than the step size in the negative direction. We will investigate specifically the branching rule that positive steps be of size Hd and negative steps be of size Ld, for $H = 2$ and $L = 3$. Also for $H = 1$, $L = 2$. Also for $H = 1$, $L = 3$. Although these are still up-and-down methods, we will also call them H-L methods. For convenience, we will speak of the up-2-down-3 method, the up-1-down-2 method, etc.

15. H-L Methods

Unless otherwise stated, subsequently reported results deal with the case where chance success occurs, with $c_i = .2$ for all items.

The H-L methods are random-walking methods designed to administer items at a difficulty level appropriate for the examinee. This should reduce the asymmetry of information curves such as those shown for $c = .2$ in Figure 5.

The chance nature of chance success introduces a random element into our measurements that necessarily must reduce their accuracy. We cannot hope to regain the lost information by tinkering with item difficulty levels.

All we can hope to accomplish is to find a better way of determining item difficulties than the simple up-and-down method of previous sections, which we can now describe as the method with $H = L = 1$.

We will continue to use the average-difficulty method of scoring. The necessary recursion equations are again (38) through (45), this time with H and L free to assume any integer values.

For fixed a and for large n , the quantity $d(H + L)/2$ is in practice roughly inversely proportional to the total number of items that will be prepared and stored in the computer (see section 20). Figure 6 compares four H-L methods, each of which has $d(H + L)/2 = .25/a$. The up-1-down-2 method and the up-1-down-3 method seem superior to the others. The up-1-down-1 method is the simple up-and-down method discussed in earlier sections. It is clearly inferior to the other H-L methods, all of which tend to favor easier items.

To avoid crowding, curves with shorter and longer average step size are not shown in the figure. On the basis of numerical comparisons, it was found that substantially reducing the step size gave poorer measurement for $\theta = \pm 2$, say, without improving matters at $\theta = 0$. Increasing the step size gave poorer measurement for $-2 \leq \theta \leq +2$ without much gain outside that range. Thus the curves shown seem to be near-optimal for the H-L methods.

16. Block Up-and-Down Methods

Let us consider other ways of modifying the simple up-and-down method so as to administer somewhat easier items in situations where answers may be correct because of chance success. If the score on a single item were polychotomous instead of dichotomous, it might be easier to arrange an up-and-down procedure under which the examinee will get about two-thirds of the

items right. This suggests combining items of equal difficulty into blocks. After a block has been administered, the score on the block is used to determine which block shall be administered next.

In bioassay, the block up-and-down method often has great advantages. For example, it may be convenient to treat several insects at once rather than one at a time. The blocking method has been investigated by Tsutakawa (1967a, 1967b, 1963), by Wetherill (1963), by Cochran and Davis (1964), and by Brownlee, Hodges, and Rosenblatt (1953). It is not clear that blocking items adds any convenience in computerized testing. It does, however, make possible more complicated branching processes than the usual random walk.

Only one block up-and-down method is investigated here. The blocks contain two items each. If the examinee answers both items at difficulty level b_i correctly, the next block will have items of difficulty $b_i + d$. If he answers only one correctly, the next two items will be at difficulty level b_i. If he answers neither correctly, the next block will be at difficulty level $b_i - 3d$.

If average-difficulty score is used, the basic recursion equation for this method is

$$X_{v+2}(b,\theta) = \begin{cases} 2b + 2d + X_v(b + d,\theta) & \text{with probability } P_t^2 \\ 2b + X_v(b,\theta) & \text{with probability } 2P_tQ_t \\ 2b - 6d + X_v(b - 3d,\theta) & \text{with probability } Q_t^2 \end{cases}, \tag{47}$$

where the random variable $X_v(b,\theta) = \sum_{r=2}^{v+1} b^{(r)}$ is the sum of the b_i values of the items administered to an examinee at ability level θ under the

specified block up-and-down method with $b^{(1)} = b$. The other necessary equations can be derived from (47) but will not be written out here.

Results of the calculations showed this particular method to be inferior to a variety of the simple H-L methods described in section 15. Many other block methods could be tried out. This has not been done here, however.

17. Plicate Methods

When $c \neq 0$, we need to produce an asymmetry, so that the examinee is more likely to give a right answer than a wrong one. An obvious device is to rule that a correctly answered item need not always be followed, as in the simple H-L methods, by a more difficult item.

Here we investigate a <u>two-ply method</u>, defined as follows. Whenever the examinee's number-right score on the items already administered is an odd number, the next item is assigned by the H-L method with $H = 1$ and $L = 1$ (this is the simple up-and-down method). Whenever the examinee's number-right score on the items already administered is an even number, we assign the next item by setting $H = 0$ and $L = 1$.

We will also investigate a <u>three-ply method</u>: when the examinee's number-right score is not a multiple of three, $H = 1$ and $L = 1$; when it is a multiple of three $H = 0$ and $L = 1$. These and other similar methods will be called <u>plicate methods</u>. The examinee's final score need not be his number-right score. Here in all cases his final score will be his average difficulty score.

The necessary basic equations for the two-ply method, again derived by the same line of reasoning used in sections 12 and 7, will be given. The extension to the three-ply and to other patterns is straightforward. The average item difficulty

$$x = \frac{1}{n} \sum_{v=2}^{n+1} b^{(v)}$$

is the examinee's score. The random variable $X(b,\theta) = \sum_{r=2}^{v+1} b^{(r)}$ is the sum of the item difficulties when the two-ply method is used with $b^{(1)} = b$, the examinee being at ability level θ.

To start with, we will need to consider the alternate two-ply method in which $H = 0$ and $L = 1$ when the number-right score is odd, and $H = 1$ and $L = 1$ when the number-right score is even. A complete set of equations will be needed for this "alternate" pattern as well as for the original pattern. A prime will be attached to all quantities computed under the alternate pattern. Thus the random variable $X'_v(b,\theta)$ is the sum $\sum_{r=2}^{v+1} b^{(r)}$ obtained by an examinee at ability level θ when the alternate pattern is used to pick the items administered.

The basic recursions are given by two formulas:

$$X_{v+1}(b,\theta) = \begin{cases} b + d + X'_v(b + d,\theta) & \text{with probability } P_t \\ b - d + X_v(b - d,\theta) & \text{with probability } Q_t \end{cases} ; \tag{48}$$

$$X'_{v+1}(b,\theta) = \begin{cases} b + X_v(b,\theta) & \text{with probability } P_t \\ b - d + X'_v(b - d,\theta) & \text{with probability } Q_t \end{cases} . \tag{49}$$

We can halve the number of equations to be written down by using the superscripts * and o with the understanding that either one of these is to be omitted while the other is to be replaced by a prime. Then, letting $H = 1$ and $H' = 0$, the following equation can represent both (48) and (49):

$$X^{o}_{v+1}(b,\theta) = \begin{cases} b + H^{o}d + X^{*}_{v}(b + H^{o}d,\theta) & \text{with probability } P_t \\ b - d + X^{o}_{v}(b - d,\theta) & \text{with probability } Q_t \end{cases} . \tag{50}$$

From this we can derive:

$$E^{o}_{v+1}(b) = E^{o}_{1}(b) + P_t E^{*}_{v}(b + H^{o}d) + Q_t E^{o}_{v}(b - d) . \tag{51}$$

$$E^{o}_{1}(b) = (H^{o}d - t)P_t - (d + t)Q_t . \tag{52}$$

$$W^{o}_{v+1}(b) = W^{o}_{1}(b) + P_t W^{*}_{v}(b + H^{o}d) + Q_t W^{o}_{v}(b - d)$$
$$+ 2P_t(H^{o}d - t)E^{*}_{v}(b + H^{o}d) - 2Q_t(d + t)E^{o}_{v}(b - d) . \tag{53}$$

$$W^{o}_{1}(b) = (H^{o}d - t)^2 P_t + (d + t)^2 Q_t . \tag{54}$$

$$D^{o}_{v+1}(b) = D^{o}_{1}(b) + P_t D^{*}_{v}(b + H^{o}d) + Q_t D^{o}_{v}(b - d)$$
$$+ [E^{o}_{v}(b - d) - E^{*}_{v}(b + H^{o}d)]\frac{\partial Q_t}{\partial \theta} . \tag{55}$$

$$D^{o}_{1}(b,\theta) = - d(1 + H^{o})\frac{\partial Q_t}{\partial \theta} . \tag{56}$$

Figure 7 compares the best results obtained with the H-L method, the two-ply method, and the three-ply method. The three-ply curve for $ad = .2$ is almost the same as the two-ply curve shown, but slightly lower. It appears that the two-ply method is slightly superior when $c = .2$ to the three-ply method. The main conclusion is that all the methods shown in the figure are almost equally good when the proper step size is used.

Figure 7 has been shaded for $\theta > 2$ and $\theta < -1$ to call attention to a way of improving measurement that has not been mentioned up to now. Given the items-- the a_i, b_i, and c_i-- the information function does not depend at all on the nature of the group tested. In particular, the information curve depends on θ and $b^{(1)}$ only through their difference $\theta - b^{(1)}$. This has been indicated in Figure 7 by labelling the base line $a(\theta - b^{(1)})$. In looking at earlier results (see discussion in section 11), we have frequently evaluated them for a group of examinees in the range $-1.5 \leq a(\theta - b^{(1)}) \leq +1.5$. Suppose we keep the same group of examinees and the same tailored testing procedure except that the first item administered is taken at difficulty level $b^{(1)} = -0.5$ instead of at $b^{(1)} = 0$. The examinees now fall in the interval $-1 \leq a(\theta - b^{(1)}) \leq +2$.

Figure 7 shows that we get better measurement in this interval than in the other one. The information function is nearly the same from one end of the interval to the other. This is the result achieved by choosing the first item administered at an easier difficulty level than we have used up to now. The two-ply method with step size $d = .5/a$ is 86 percent efficient at $\theta - b^{(1)} = .5/a$ compared to the standard test. That is, it produces as much information with $n = 60$ items as would a standard test, $b_i = 0$, with $n = 52$ items. Here we have assumed, as in the last several sections, that all icc are normal ogives and that all $c_i = 0.2$. Except at $\theta - b^{(1)} = .5/a$ the two-ply test is more than 86 percent efficient compared to the standard test. At $\theta - b^{(1)} = -1/a$ the standard test is only 49 percent efficient compared to the two-ply test: At $\theta - b^{(1)} = +2/a$, the standard test is only 30 percent as efficient.

The standard test is a "peaked" test with all items of equal difficulty, $b_i = 0$. This produces unusually accurate measurement around one value of θ (near $\theta - b^{(1)} = 0.3/a$ in our case) at the expense of less accurate measurement at extreme values of θ . The typical published test has a wide range of b_i values (partly because it is easier to produce such a test and partly because most people believe a range of item difficulty is necessary to secure good measurement. Actually, for typical groups and typical values of a_i , the peaked test would usually provide better measurement for all except the most extreme examinees in the group tested, as is illustrated in Figure 7). Actual b_i values, estimated for a well-known published test, were used to compute the information curve labelled "published test" in Figure 7. The curve shown was computed from Birnbaum's equation (1968, eq. 20.2.2)

$$I(\theta,x) = \frac{[\sum_{i=1}^{n} P_i'(\theta)]^2}{\sum_{i=1}^{n} P_i(\theta)Q_i(\theta)}$$

under the assumption that a_i was the same for all items and also that $c_i = .20$ (the actual published test does not satisfy this assumption). The 60 values of b_i ranged from -2.5 to $+1.5$. We note that the tailored tests all measure more accurately at all values of θ than does such an unpeaked test.

18. Item Economy

The theory up to this point has assumed that items were available at whatever difficulty level was required by the random walk procedure. In

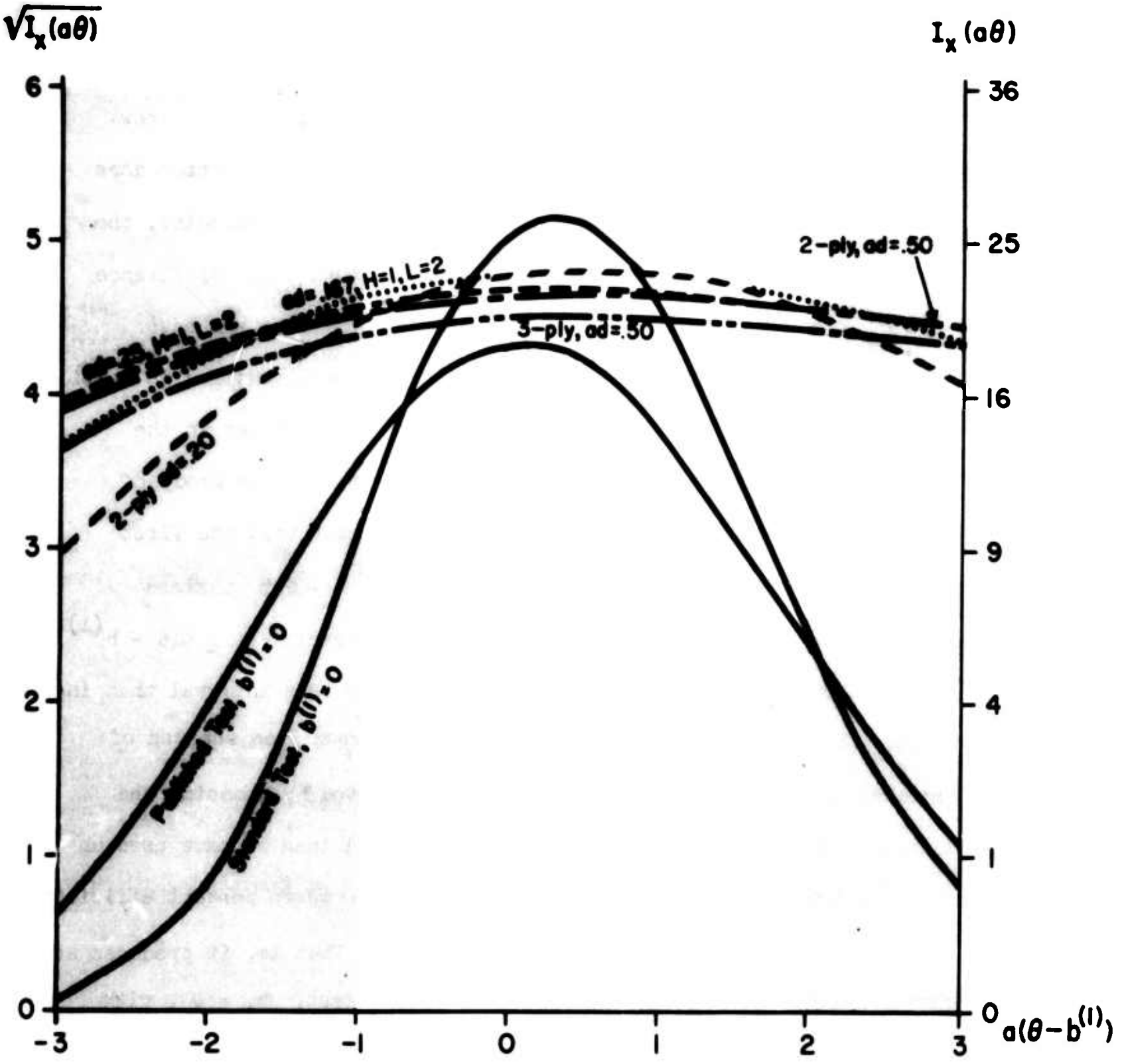


Fig. 7. A comparison of H-L methods with plicate methods.

principle, this would require for the H-L or plicate methods a total of $n(n + 1)/2$ items, regardless of step size. When $c = 0$, there is usually a vanishingly small probability of needing items with $b_i > 5$, say. If we supply all the items that theoretically might be required in the range $-5 \leq b_i \leq 5$ and no items outside this range, the simple up-and-down method will require, depending on the details of the random walk, about $(1 + 5/d)(n - 5/2d)$ items, assuming d is a submultiple of 5. For sizable n, this quantity is roughly inversely proportional to the step size d. For $n = 60$, $d = .50$, the number of items required would be about 600.

Presumably this number could be considerably reduced without much loss of accuracy of measurement. Whenever items are in short supply (many shortages should disappear, given enough time) practical applications of tailored testing will have to consider which short-cuts will cause the least loss of efficiency. Much light could be thrown on this question by Monte Carlo studies of different possible procedures.

It would be desirable to assign a cost to items, a cost to computer storage space, and a loss function to errors of measurement. One could then investigate the economic efficiency of various tailored testing procedures under various cost conditions. Something of this sort will have to be done, whether formally or informally, before tailored testing comes into widespread use. Nothing like this has been attempted for the present report.

19. Robbins-Monro Processes

It will have occurred to the reader that a large step size is needed for the first few items so that the item difficulty can be rapidly adjusted to suit an extreme examinee. Once this has been done, progressively smaller step sizes are needed in order to zero in on the best value.

Probably considerable gain could be achieved simply by choosing $d = 1$, say, for the first three or four items and thereafter proceeding with a smaller but fixed d, by any of the methods of the preceding sections. This rule has not yet been investigated for tailored testing.

Carrying this idea to its logical conclusion leads to a Robbins-Monro type of process. Let $P(b,\theta)$ denote any monotonic increasing item characteristic curve satisfying certain regularity assumptions. An important advantage of the method is that we do not have to know the mathematical form of $P(b,\theta)$. We wish to find the value of b for which $P(b,\theta)$ equals some chosen constant, α (under the normal-ogive model of equation 3, α is usually chosen as $1/2$, since $P(\theta,\theta) = 1/2$).

In a Robbins-Monro process, we choose a decreasing sequence of positive constants $A_1, A_2, A_3, \ldots$, to be discussed later. We choose a trial value of b for the first item— let this be $b^{(1)} = 0$. We administer the item and observe the response $u_1 = 0$ or 1. Then we choose all subsequent values of $b^{(v)}$ by the rule

$$b^{(v+1)} = b^{(v)} + A_v(u_v - \alpha) . \tag{57}$$

This leads to an H-L method with continually shrinking step size. Robbins and Monro (1951) showed that under suitable conditions $b^{(v)}$ converges in probability to θ as $v \to \infty$, that is, $b^{(v)}$ <u>is a consistent estimator of</u> θ.

Chung (1954) and Hodges and Lehmann (1956) showed that under certain conditions, satisfied for our problem, if the A_v are determined from

$$A_v = C/v \quad , \tag{58}$$

where C is a constant to be determined, the asymptotic sampling variance of $b^{(v)}$ will be minimized. Hodges and Lehmann (1956, section 3) show that for our kind of quantal response problem, the value of C minimizing the asymptotic sampling variance is

$$C = - \frac{1}{\partial P(b,\theta)/\partial b \Big|_{b=\theta}} \quad . \tag{59}$$

Of course, this value cannot be determined without knowing the nature of $P(b,\theta)$. In our case, this optimal value for the normal ogive model of (5) is

$$C = \frac{\sqrt{2\pi}}{a_i(1 - c_i)} \quad . \tag{60}$$

If $c_i = 0.2$ and $a_i = 1.0$, then the initial step size under (58) and (60) is $A_1 = C = 3.1$. If $a_i = .333$, then $A_1 = 9.4$, if $a_i = .10$, then $A_1 = 31.3$, etc.

However, we have been assuming that almost all our examinees fall within some roughly known interval centered about the origin of our θ scale. If we have such information (obtained from previous testings) and if we are willing to forego accurate measurement of the very rare examinee who falls outside the interval, then we probably do not need to use an initial step size greater than about half the width of the interval.

It would be very desirable to obtain information functions for the test scores $b^{(n)}$ produced by the stochastic approximation method outlined.

Unfortunately, there appears to be no available method of doing this precisely for a test longer than a dozen or so items. Recursion formulas cannot be worked out by the methods used in earlier sections.

Asymptotic formulas are available, but these would produce information functions that are horizontal straight lines. Previous sections indicate that optimum step size is found by reducing step size as far as possible without making the information curve fall off too sharply on both sides of the maximum. Asymptotic results would be useless for this purpose.

The only good way to deal with this problem would seem to be by using Monte Carlo procedures. For various reasons, we have not used any Monte Carlo procedures for this study. Wetherill (1963) made extensive Monte Carlo investigations of the Robbins-Monro method from the bioassay point of view. He found that the method was not sensitive to mischoice of C. The method was "extremely satisfactory" in the case where $P_i(b^{(n)},\theta) \to \frac{1}{2}$ as n becomes large (we have this case when $c_i = 0$); but "of little use," "unsuitable," "hopeless" in the case where $P_i(b^{(n)},\theta) \to .75$ (we have this case when $c_i = .5$).

The main trouble in this latter case ($P_i \to .75$) is the large bias produced by the Robbins-Monro method. As already pointed out in section 11, bias is a serious matter in bioassay work, but usually of no concern in mental testing. For this reason, and for other reasons indicated in section 11, it is difficult to apply many of Wetherill's reported results to the purposes of the present study. [Wetherill also tried out many promising up-and-down methods, which could not be evaluated by the techniques used for this study. These could and should be investigated for tailored testing purposes by use of Monte Carlo methods.]

The use of (58) to decide on step size is practical in bioassay, but is not practical in tailored testing. In principle, use of (58) would require that almost 2^n different items be available to the computer--an impossible requirement if $n = 60$. An obvious modification is to classify all items on b_i into class intervals of width d, and then pick a sequence of class intervals corresponding as closely as possible to the sequence specified by (57).

This modification destroys all the asymptotic virtue of the Robbins-Monro process, since the use of fixed class intervals prevents convergence of $b^{(n)}$ to θ as $n \to \infty$. However, this difficulty need not impair the information actually produced by a test of fixed length. If a psychometrician is going to administer a test of fixed length n, there is no good reason why he should use a consistent estimator of θ.

The way of choosing items outlined above was tried out for this study for short tests with $n = 10$ and $c_i = .20$. With $n = 10$, it was possible for the computer to deal individually with each of the 2^{10} possible patterns of examinee response. Average difficulty score was used, not final difficulty score, as in the standard Robbins-Monro process.

No choice of C was found that yielded information curves as good as those obtained by the better up-and-down methods. Detailed results will not be given here.

Another possibility for improving the estimation of θ is to use a weighted rather than an unweighted average of item difficulties for the examinee's score. The difficulty $b^{(1)}$ of the first item has always been omitted from the average difficulty score (19) since it is the same for all examinees and thus cannot carry any information about θ. If step size is small, the same objection applies to a lesser extent to $b^{(2)}$. Clearly, the item difficulties for the earlier items are not expected to

be as close to θ as those for the later items. This suggests that the later items should receive more weight in determining final score than the earlier items.

Two systems of weighting were tried out here:

$$x = \sum_{v=2}^{n+1} \sqrt{v}\ b^{(v)}\ , \tag{61}$$

$$x = \sum_{v=2}^{n+1} v b^{(v)}\ . \tag{62}$$

The weighted score (61) with weights proportional to the square root of the item serial number v was found to be a little better than the unweighted average difficulty score. The weighted score (62) with weights proportional to item serial number was still better than (61).

It would be desirable to have results for $n = 60$ as well as for $n = 10$. In view of the incomplete nature of presently available results, no information functions will be displayed here. There is clearly a need for extensive Monte Carlo studies to investigate further the methods already outlined and the numerous possible recombinations and mutations of these methods.

20. Summary and Conclusions

When computers are used extensively for instruction, it will be convenient to use them to administer tests for measurement purposes as well as for instructional purposes. We restrict attention here to tests used for measurement. Given a supply of pretested items, the computer can individually design a test for each individual examinee. How can it secure the most accurate measurement?

The question falls into two parts: What items shall be administered to a particular examinee? How shall his responses be scored? We do not find optimum procedures. We compare various simple procedures and answer a number of primitive but previously unanswered questions.

A number of empirical studies of tailored tests have been carried out using an external criterion to evaluate the results. Here we have no external criterion. Our purpose is to measure, not to predict.

The examinee can be measured by the following simple <u>branching rule</u>: Administer a harder item after each correct answer, an easier item after each wrong answer. Start with large changes in item difficulty (b), then use smaller and smaller steps, so as to zero in on the difficulty level at which the examinee answers correctly 50 percent (say) of the time. This final difficulty level, $b^{(n+1)}$, is a measure of the examinee's standing on the trait measured by the test.

This is the Robbins-Monro stochastic approximation process. Asymptotically optimum details for the procedure are known. To investigate its efficiency in subasymptotic situations requires Monte Carlo methods. We do not use these methods here.

Instead, we investigate various branching rules that do not shrink the step size as testing progresses. We also investigate three different scoring methods. We evaluate the efficiency of these procedures for short, medium, or long tests, as desired. Our problems are found to be very closely related to certain problems in bioassay. Many important results obtained for the bioassay problems are of direct use to us here.

To keep matters simple, we assume that available items differ only in difficulty (b_i). In order to compare the efficiency of various branching and scoring procedures, we often assume that the probability of

a correct answer to an item is a normal ogive function of θ, the examinee's standing on the trait measured. Alternatively, a logistic function is assumed. Both assumptions are generalized to cover the case where correct answers may be due to guessing.

Conventional tests ordinarily provide good measurement for the middle three-quarters, say, of the group tested. The tailored test cannot hope to provide much improved measurement for these examinees, but it can provide better measurement at higher and lower levels of θ. In view of this picture, we are not satisfied to use an overall group statistic to describe the effectiveness of measurement; instead, we use an information function to tell us the accuracy of measurement at each level of θ.

We compute and compare information curves for a variety of procedures. Some of the conclusions tentatively reached for a certain specified, presumably typical,* tailored test using a simple up-and-down branching rule are listed below. This rule increases (decreases) item difficulty by an amount d after each correct (incorrect) response.

1. If the test score is $b^{(n+1)}$ and the step size d is .50, increasing the number of test items beyond $n = 10$ does not appreciably increase the accuracy of measurement.
2. The number of items answered correctly is perfectly correlated with the score $b^{(n+1)}$.
3. When the foregoing scores are used, decreasing the step size increasec the accuracy of measurement near some one value of θ and decreases it elsewhere.

---

*The conclusions given are for items with $a_i = .5$.

4. The average difficulty of all items administered is a score providing better measurement of the examinee than either $b^{(n+1)}$ or the number of right answers.

In the conclusions that follow, the score used is always average difficulty.

5. For $n = 60$, a step size of around .40 seems best; for $n = 10$, a step size of around 1.0. Either shortening, or lengthening the step size decreases the accuracy of measurement throughout the range of θ that interests us.

6. When items can be answered correctly by chance success, the accuracy of measurement is sharply reduced. Also, the information curves become asymmetrical.

7. When there is chance success, the accuracy of measurement can be considerably increased by certain asymmetric modifications of the step size used in the simple up-and-down method. Two such modifications are found to be almost equally good.

8. When there is chance success, the first item administered should be easier than that specified by a common rule of thumb.

9. These improvements produce nearly symmetrical and reasonably flat information curves.

10. When these tailored testing procedures are compared with a "standard" conventional peaked test, also with a conventional unpeaked test, both scored by the number of right answers, we see that the best tailored testing procedure is nowhere less than 86 percent efficient compared to the peaked test. For high-level examinees, the peaked test is only about 30 percent efficient compared to the tailored testing. The tailored procedure gives more accurate measurement than the unpeaked conventional test for all examinees, regardless of level.

Before closing, let us note some of the limitations of tailored testing procedures and of the theory given here.

1. Suppose a pool of test items can be grouped into subtests measuring substantially different psychological dimensions. Without such grouping into subtests, tailored testing based on such a pool cannot produce accurate measurements with a clear meaning.
2. The theory given here assumes that items differ from each other only on difficulty level. In practice, they differ also on a_i (discriminating power) and on c_i (see text). It is an open theoretical question how tailored testing should be modified to deal with this more general situation.
3. Accurate estimation of the item parameters necessary for tailored testing is at present a difficult, expensive, and hazardous operation.
4. If there is any doubt about the accuracy of the estimated values of the item difficulties, b_i, there will be doubt about the accuracy and fairness of the final scores given to the examinees.
5. If, say, 500 items are available for tailored testing, better measurement will often be obtained by selecting, say, the $n = 60$ most discriminating items (highest a_i) and administering these as a conventional test, rather than by using all 500 in a tailored testing procedure.

Until now, even some very primitive questions about how to carry out tailored testing did not have even vague answers. Granted certain assumptions, we now have tentative answers to some of these questions. More important, we have a theoretical approach, drawing heavily on bioassay theory and results, and on the theory of stochastic processes with particular reference

to Markov chains. This theory shows how we can go ahead to evaluate the endless variety of different possible combinations of branching processes and scoring procedures available for tailored testing. Perhaps in due course, some direct way of finding truly optimum tailored testing strategies will be found.

The theory in this report is based on certain rather technical and specialized assumptions. Most of the conclusions reached are, hopefully, of more general validity. The tailored testing procedures themselves can provide accurate measurements without any need for many of these assumptions.

We have investigated in detail only those tailored testing procedures that could be rigorously evaluated numerically for tests of 60 items (and longer). Up-and-down item-selection methods with continually shrinking step size (section 19) should be able to produce more accurate measurement than is obtained by methods without a shrinking step size. There is a clear need for studies to evaluate various possible shrinking-step-size procedures for tests much longer than $n = 10$ or 12. These studies will probably have to be carried out by Monte Carlo methods.

Small sample properties of the maximum likelihood estimator of θ (Tsutakawa, 1967a, 1967b; Billingsley, 1961; Dixon, 1965; Dixon & Mood, 1948) should also be investigated. Another estimation method requiring further study by Monte Carlo methods is the Spearman-Kärber method (Spearman, 1908; Kärber, 1931). This method is described and favorably evaluated for bioassay purposes by Tsutakawa (1967a).

APPENDIX

This appendix outlines one or two results from the theory of Markov chains, relevant for evaluating tailored testing procedures.

We are concerned (as in section 5) with the random variable $b^{(v)}$, the difficulty of the v-th item administered to a given examinee. The possible values of $b^{(v)}$ are (see eq. 15) $b^{(v)} = jd$, where d is some prespecified step size and j is a (possibly negative) integer with $|j| \leq n$. There will be no loss of generality for the purposes of this appendix if we rescale $b^{(v)}$ so as to set $d = 1$, in which case $b^{(v)}$ only takes on integer values between $-n$ and $+n$, inclusive. Denote such integer values by either j, k, or i.

Define the <u>transition probability</u>

$$p_{ij} = \text{Prob}\ (b^{(v+1)} = j | b^{(v)} = i)\ , \quad v = 1,2,\ldots\ . \tag{63}$$

By the Markov property, this does not depend on $b^{(1)},\ldots,b^{(v-1)}$. We will consider only <u>stationary transition probabilities</u>, which means that p_{ij} does not vary with v (this rules out all branching methods with shrinking step size).

Define the <u>r-step transition probabilities</u>

$$p_{ij}^{(r)} = \text{Prob}\ (b^{(v+r)} = j | b^{(v)} = i)\ , \quad r,v = 1,2,\ldots\ . \tag{64}$$

It is easily seen that

$$p_{ij}^{(2)} = \sum_k p_{ik}p_{kj}\ .$$

If the $p_{ij} = p_{ij}^{(1)}$ are written as the elements of a square matrix $P = \|p_{ij}\|$ of order $2n + 1$, then the $p_{ij}^{(2)}$ are elements of $P^2 = PP$. Similarly, the $p_{ij}^{(r)}$ are elements of P^r:

$$\|p_{ij}^{(r)}\| = \|p_{ij}\|^r . \tag{65}$$

For any given examinee, the nonzero elements of the matrix P are the $P_i(\theta)$ and $Q_i(\theta)$ of (1), (2), (3), or (5), or simple functions of these, depending on the branching process chosen.

Let p_{oi} denote the probability that $b^{(1)} = i$ and let p be the vector $\{p_{oi}\}$. If we choose our origin so that $b^{(1)} = 0$, then p_{oi} is zero, except that when $i = 0$, then p_{oi} is 1.

The final frequency distribution of $b^{(n+1)}$ for a given examinee is thus the vector $P'^n p$:

$$\{\text{Prob}\ (b^{(n)} = i|\theta)\} = P'^n p . \tag{66}$$

The mean and variance of this distribution are important quantities related in a simple way to those computed recursively by (25) and (30). These quantities could be computed directly from $p'P^n$. The matrix P^n may be computed from the latent roots and vectors of PP' and of $P'P$ (Feller, 1959, chapt. 16, eq. 1.12).

As already noted in section 19, asymptotic results are of marginal interest for present purposes. Asymptotic properties of $b^{(n+1)}$ can be found from Chung (1960, part 1, section 12). Although the $b^{(v)}$ form a Markov chain, the average difficulty scores do not. The average difficulty score is a functional of the Markov chain. Asymptotic properties of such functionals are treated by Chung (part 1, sections 14-16). An asymptotic formula for the error variance of the average difficulty score is given by Tsutakawa (1967a, eq. 5).

References

Billingsley, P. Statistical inference for Markov processes. Statistical Research Monographs, Volume II. Chicago, Ill.: The University of Chicago Press, 1961.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968. Chapters 17-20.

Bock, R. D. Fitting a response model for n dichotomous items. Paper read at the Psychometric Society Meeting. Madison, Wisconsin, 1967.

Brownlee, K. A., Hodges, J. L., Jr., & Rosenblatt, M. The up-and-down method with small samples. Journal of the American Statistical Association, 1953, 48, 262-277.

Chung, K. L. On a stochastic approximation method. The Annals of Mathematical Statistics, 1954, 25, 463-483.

Chung, K. L. Markov chains with stationary transition probabilities. Germany: Springer-Verlag, 1960.

Cochran, W. G. & Davis, M. Stochastic approximation to the median effective dose in bioassay. In J. Gurland (Ed.) Stochastic models in medicine and biology. Madison, Wisc.: University of Wisconsin Press, 1964.

Dixon, W. J. The up-and-down method for small samples. Journal of the American Statistical Association, 1965, 60, 967-978.

Dixon, W. J. & Mood, A. M. A method for obtaining and analyzing sensitivity data. Journal of the American Statistical Association, 1948, 43, 109-126.

Feller, W. An introduction to probability theory and its applications. Volume I, 2nd Edition. New York: John Wiley & Sons, Inc., 1959.

Hansen, D. N. & Schwarz, G. An investigation of computer-based science testing. Tallahassee, Florida: Institute of Human Learning, Florida State University, 1968.

Hodges, J. L., Jr. & Lehmann, E. L. Two approximations to the Robbins-Monro process. In J. Neyman (Ed.), Proceedings of the third Berkeley symposium on mathematical statistics and probability. Volume 1. Calif: University of California Press, 1956.

Kärber, G. Beitrag zur kollektiven Behandlung pharmakologischer Reihenversuche. Archiv fur experimentelle Pathologie und Pharmakologie, 1931, 162, 480-487.

Linn, R. L., Rock, D. A., & Cleary, T. Anne. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969 (in press).

Lord, F. M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lord, F. M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-76.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968 (in press). (a)

Lord, F. M. Estimating item characteristic curves without knowledge of their mathematical form. Research Bulletin 68-8 and ONR Technical Report, Contract Nonr 2752(00). Princeton, New Jersey: Educational Testing Service, 1968. (b)

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley Publishing Company, 1968.

Mandel, J. & Stiehler, R. D. Sensitivity--A criterion for the comparison of methods of test. Journal of Research of the National Bureau of Standards, 1954, 53, 155-159.

Robbins, H. & Monro, S. A stochastic approximation method. The Annals of Mathematical Statistics, 1951, 22, 400-407.

Spearman, C. The method of "right and wrong cases" ("constant stimuli") without Gauss' formulae. British Journal of Psychology, 1908, 2, 227-242.

Tsutakawa, R. K. Block up-and-down method in bio-assay. Doctoral dissertation, University of Chicago, 1963.

Tsutakawa, R. K. Random walk design in bio-assay. Journal of the American Statistical Association, 1967, 62, 842-856. (a)

Tsutakawa, R. K. Asymptotic properties of the block up-and-down method in bio-assay. The Annals of Mathematical Statistics, 1967, 38, 1822-1828. (b)

Turnbull, W. W. Relevance in testing. Science, 1968, 160, 1424-1429.

Wetherill, G. B. Sequential estimation of quantal response curves. Journal of the Royal Statistical Society, 1963, 25, 1-38.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

Educational Testing Service
Princeton, New Jersey 08540

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

SOME TEST THEORY FOR TAILORED TESTING

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Technical Report

5. AUTHOR(S) *(First name, middle initial, last name)*

Frederic M. Lord

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
September 1968	62	9

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Nonr 2752(00)	RB-68-38
b. PROJECT NO. NR 151-201	
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research Navy Department Washington, D. C. 20360

13. ABSTRACT

In a tailored test, each item is selected for administration on the basis of the examinee's responses to previous items, with a view towards optimum measurement of this particular examinee. Various simple rules for 1) selecting the items to be administered and 2) scoring the examinee's responses are compared and evaluated. Some fundamental ideas emerge that will serve as guides in the future design of tailored testing programs.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

Security Classification

3ND FP30 13152

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
mental test theory						
Markov chains						
stochastic approximation						
computer applications						
computer-assisted instruction						
bioassay						
random walks						